Statistical Analysis of Genome-Genome Interaction with Reference to Kidney Transplant Outcome

Mollon, Jennifer

Awarding institution:
King's College London

Statistical Analysis of Genome-Genome Interaction with Reference to Kidney Transplant Outcome

Jennifer Mollon

A dissertation submitted to King's College London for the degree of

Doctor of Philosophy

MRC Centre for Transplantation

Guy's Hospital

Great Maze Pond

London SE1 9RT

United Kingdom

October 2012

ABSTRACT

Though widely believed to exist, few convincing examples of genetic interactions have been detected through statistical approaches in genome-wide association studies. The first piece of work in this thesis attempts to determine if there is evidence for the existence of such interactions within genes identified through protein-protein interactions. A software package is developed and applied to data from a recent publically available genetic study. No evidence was found for an enrichment of such interactions in the available data.

The second study applies three current methods for interaction detection to a real data set with compelling evidence of an interaction. Sparse Partitioning, SNPHarvester and Random Jungle were selected, with the later two being followed by the HyperLasso as a post-processing step. Only one method, SNPHarvester, was able to detect the interaction.

The third study outlines a local pilot project in renal transplant dysfunction. Genetic variants from donors and recipients are examined using survival analysis. Interactions between the two genomes are tested for an effect on the survival time of the kidney. Secondary renal phenotypes of acute rejection and progression to end-stage renal failure are also considered. There were no strongly significant associations discovered in this data.

The final study is a multi-centre renal transplant study analysing over 2000 donor recipient pairs throughout the UK and Ireland. Although much larger than the pilot, this study also failed to detect any associations with graft survival time or the secondary phenotypes. SNPHarvester was applied to the data and there are some indications of potential interactions, but replication is essential before the results can

be trusted. An outline of an extension to SNPHarvester to better handle survival data is presented.

Results from all of these studies were largely negative. Detecting interactions in genome-wide data remains a difficult task. Narrowing the search space by filtering may be a better approach than attempting a genome-wide search, though SNPHarvester has proven to be useful and is a good choice if a true genome-wide search is required.

Acknowledgements

First and foremost I am indebted to my supervisor Mike Weale for his support, patience and kindness. His experience and guidance has been invaluable.

I am grateful to the staff and students, past and present, of the statistical genetics group, Cathryn Lewis, Daniel Crouch, Sarah Spain, Graham Goddard, Jo Knight, Paola Forabosco, Adai Ramasamy and Ian Scott, for generously sharing your knowledge, time and skills with me. Thank you to the bioinformatics team, Thomas Schlitt, Benni Lehne, Nick Dand, Russell Sutherland, for insights, interesting conversations and collaborations.

To the team in the MRC Centre for Transplantation, I thank you for your guidance and assistance through the course of my studies. I would particularly like to acknowledge my second supervisor Steve Sacks, Graham Lord, Florence Delaney, Espe Perucha, Jill Holliday, Amii Adams and Stephanie Goldberg.

To my parents Anne and Paul Mollon, who taught me to love reading, learning and debating, and to always ask 'why'.

Finally my partner in life, Richard. I thank him for being my inspiration to work towards a PhD and for his supportive presence through the good times and the bad. I couldn't have done it without him.

Table of Contents

Table of Figures

Table of Tables

Table of Equations

1 Introduction

In recent years the field of genetics has benefitted from significant advances in technology. Measuring a person's genetic makeup is becoming increasingly faster and cheaper, however an understanding of how this variability affects us and makes us who we are is more difficult to achieve. Many hundreds of thousands, or even millions, of pieces of genetic information can be available for every individual. Making sense of the vast amount of genetic information is a daunting task. Additional non-genetic measurements and an ever-increasing number of possible outcomes complicate the picture further. Simple statistical approaches have met with some success but a deeper understanding of the complex genetics behind many human traits may be possible with further methodological development. This is particularly true in the field of renal transplantation where, as I will explain later, the success of the transplant can be viewed as a complex trait with important implications for the health of the recipient.

1.1 Human Genetics

The human genome contains the genetic material that, combined with the environment in which we develop and live, determines our physical characteristics. Either directly or in combination with environmental factors, it determines our sex, eye colour, hair colour, height and many other physical characteristics, as well as our predisposition to certain medical conditions and diseases. The way in which our genome is read and interpreted, and its instructions carried out, is summed up concisely in what is called the central dogma of molecular genetics.

1.1.1 Proteins and the Central Dogma of Molecular Biology

Proteins are essential to life. They make up about 45% of the mass of a human, and are present in skin, hair, muscles, and just about any structure within the human body. Proteins act as enzymes, are involved in communications between cells ('cell signalling'), and act as antibodies, a critical part of the immune system that keeps us healthy. In fact, most basic cellular processes could not occur without proteins. Some proteins come from the food we eat, but many are produced by the body.

In 1958 Francis Crick first stated the 'Central Dogma' of genetics:

This states that once 'information' has passed into protein it cannot get out again. In more detail, the transfer of information from nucleic acid to nucleic acid, or from nucleic acid to protein may be possible, but transfer from protein to protein, or from protein to nucleic acid is impossible. Information means here the precise determination of sequence, either of bases in the nucleic acid or of amino acid residues in the protein.

(Crick, 1958)

Crick restated this in Nature in 1970:

The central dogma of molecular biology deals with the detailed residue-by-residue transfer of sequential information. It states that such information cannot be transferred back from protein to either protein or nucleic acid.

(Crick, 1970)

It is often put somewhat more simply:

"DNA makes RNA makes protein."

To understand the central dogma, it is important to understand proteins, DNA, RNA, and why they are important.

A protein is formed from a linear chain of amino acids which then folds into a final three-dimensional shape. There are 12 amino acids synthesized by the human body, and these are joined up in various sequences to form different proteins. The sequence of amino acids dictates which protein is produced following processes called transcription and translation.

1.1.2 Genes, transcription and translation

All humans (and most living organisms) carry their genetic makeup in a molecule called deoxyribonucleic acid (DNA). DNA exists in long strands in the nucleus of almost every cell in the body. A single strand of DNA is a polymer; this means it is a molecule that is formed by many similar molecular units bonded together linearly. The smaller molecular units, or monomers, are called nucleotides. There are four nucleotides that make up DNA; cytosine, guanine, adenine and thymine. A unit of three nucleotides bonded together in sequence is called a codon. Most codons encode a specific amino acid, or indicate the beginning or end of the coding sequence of a gene (start/stop codons). A gene is considered to be a basic unit of heredity of an individual, and each one contains the sequence of bases that dictate the amino acid sequence that makes up a single protein.

Molecules of ribonucleic acid (RNA) are formed in the nucleus of a cell during a process called transcription. RNA molecules share many of the properties of DNA. A strand of RNA is also a sequence of nucleotides bonded together in a chain. In RNA, however, the nucleotide thymine is not present; instead it is replaced by uracil. Another difference is that DNA is double-stranded - two strands align together and

the bases bond together in a specific manner to form the double-helix ladder-like structure of a DNA molecule. Between the two strands, cytosine will bond only with guanine and adenine only with thymine. Because of this the two strands are 'compliments' of each other, and the sequence of one strand is predicted entirely by the other strand. During transcription an enzyme called RNA polymerase separates the two strands, reads the sequence on one strand, and creates a complimentary strand of RNA from this template. The sequence of bases on the RNA molecule is completely complementary to the sequence on the DNA molecule except that where a thymine is expected, it is replaced with a uracil on the RNA strand.

After a strand of RNA is produced, sections of it are removed ('spliced') in a process called post-translational modification. Non-coding sequences can exist between genes but also inside them, and the non-coding sequence must be spliced out of the RNA so that only the correct amino acids are include in the protein. A region of protein-encoding sequence within a gene is called an exon; a non-coding region within a gene is called an intron.

Translation is the process by which a cell produces a protein using RNA. The RNA, after the non-coded regions have been spliced out, acts as a template for the amino acid sequence. A cellular complex called the ribosome reads the template and attaches amino acids in the correct sequence, indicated by the codons, to form the protein.

1.1.3 Chromosomes and the Size of the Human Genome

Each haploid set of chromosomes in a human contains 3 billion bases, for a total diploid set of 6 billion bases. There are approximately 23 000 protein-encoding genes, and only about 1.5% of the human genome is actually protein-coding

sequence. The exact function of the rest of the DNA is not known exactly, but some regions play a role in regulating how much protein is produced. Genes are not always being transcribed and translated; a gene whose protein product is currently being produced is said to be 'expressing'. Regulatory regions can control the expression levels of a gene.

DNA is folded and wrapped around structural proteins called histones to form a very dense structure called a chromosome. Chromosomes are located in the nucleus of a cell. The normal human genome is made up of 2 copies of each of 23 chromosomes. One pair of chromosomes determines the sex of the individual; females have two copies of the X chromosome, while males have one X and one Y chromosome. The other 22 pairs of chromosomes are referred to as 'autosomal' chromosomes. Each parent of an individual donates a single chromosome towards each of the 23 pairs, hence 50% of the genetic material is inherited from the mother and 50% from the father.

1.1.4 Recombination and Linkage Disequilibrium

A chromosome inherited from a parent is created from a combination of their two chromosomes. The genetic material of the two chromosomes is combined in a process called recombination. On average, each of a parent's chromosomes contributes 50% of their genetic material to the chromosome inherited by the offspring. The points at which chromosomes cross over and recombine are not random; some places are for more likely to have a crossover event than others, leading to areas called recombination hotspots. Between these hotspots, chunks of DNA are likely to be inherited together, leading to highly correlated sequences in these regions. Correlation between polymorphisms due to physical proximity on a

chromosome is called linkage disequilibrium (LD). Between any two loci on the same chromosome this can be measured by calculating the square of the correlation coefficient that would be obtained by coding alleles as 0 or 1 and observing their co-occurrence. This measure is referred to as r^2, and requires that the sequence of variants on each of an individual's chromosomes is known. Assigning variants to each chromosome can be done in a process called phasing. R^2 is often more conveniently estimated without resorting to phasing by coding each locus as carrying 0,1 or 2 copies of an allele and observing the within-individual co-occurrence of these values. An r^2 of 1 indicates perfect correlation between two loci, and the allele at one locus can always be perfectly predicted by the allele at another. An r^2 of 0 indicates a completely random relationship between the two loci. This local LD is crucial to the success of genome-wide association studies, as it allows a single variant on a GWAS panel to represent or 'tag' a large number of common variants around it.

1.2 History of genetic association studies

1.2.1 Discovering molecular markers

A molecular marker for a genetic study is a fragment of DNA that is measurable, identifiable, and varies between individuals. These markers can be as small as single nucleotide or involve longer lengths of DNA sequence. The marker itself may not directly influence a trait or diseases, but will 'tag' regions of the genome that differ between individuals. The underlying assumption is that people with the same genetic markers will have inherited the same DNA sequence near that marker, including variants that are not directly measured.

Early genetic studies used markers called restriction fragment length polymorphisms (RFLPs) in tests of associations. DNA is cleaved at specific locations called restriction sites. Variability in the length of the resulting DNA fragments, as measured by techniques such as the Southern blot, indicate heritable changes in the DNA sequence between restriction sites. These can result from several types of genetic variation such as insertions or deletions of DNA, variable number tandem repeats (VNTRs), translocations, or changes in the sequence of the restriction site itself. These polymorphisms can be used in genetic analyses to identify regions of the genome associated with traits or diseases. RFLP analysis is useful but the process is slow and labour-intensive.

The discovery of polymerase chain reaction (PCR) methods to amplify DNA simplified the discovery of molecular markers. Oligonucleotides called primers are designed to match sequences flanking the region of DNA to be amplified. The DNA sequence between these sites is replicated in a series of heating and cooling cycles. Replication is exponential since the newly-generated DNA is amplified along with the original DNA in subsequent cycles. The PCR products can then be run through a gel which separates fragments of different lengths. In genetic studies PCR was commonly used to identify microsatellites, which are short sequences (2-6 base pairs) of nucleotides that repeat. The number of repeats and the length of the microsatellites vary between individuals, making them useful markers.

In 1977 Fred Sanger developed a method now referred to as Sanger sequencing. "Sequencing" refers to techniques which determine the exact sequence of nucleotides

in a strand of DNA. In Sanger's method, single-stranded DNA is replicated in vitro in four separate tubes, each one containing all 4 nucleotides plus small amounts of one of four "ddNTP" molecules. ddNTPs are the same as nucleotides but they prevent the bonding of further nucleotides and terminate the DNA strand. Thus, all strands in the "G" tube are known to terminate with guanine. However the strands will be different lengths, because the ddNTP will be incorporated randomly at any position where guanine can bond. The products from all four tubes are then run side-by-side on a gel which separates the DNA strands by length. The order of the strands by fragment lengths indicates the position in the sequence, and the nucleotide at that position is determined by the tube in which the fragment was produced. The development of Sanger sequencing was a large step forward in the field of genomics, though it has now been largely replaced by the introduction of high-throughput parallel sequencing methods such as Solexa sequencing and 454 pyrosequencing. However it is still considered to be a 'gold standard' method, and it is still used to verify the results of other techniques and for small sequencing projects.

1.2.2 Mapping the genome

The development of Sanger sequencing made it possible to consider reading the DNA sequence of large regions of the genome. Beginning in 1990, the Human Genome Project aimed to read not just selected regions, but the entire DNA sequence of a human genome. The first draft of the human genome appeared 10 years later (Lander et al, 2001), and the final draft was published three years later. Further reference genomes have been sequenced and published since, notably through the

HapMap project (International HapMap Consortium, 2003) and the 1000 Genomes project (1000 Genomes Project Consortium, 2010). Projects such as these have enabled the creation of a 'reference genome' using samples from multiple individuals. The reference genome identifies nucleotides present at every locus, including information on allele probabilities at polymorphic sites across different ethnic groups. Data from all of these studies is freely available to researchers for use in their own genetic studies.

Information gathered from projects such as the Human Genome Project and the HapMap project made it possible to identify much of the variation across the human genome. This information allows researchers to design techniques to measure genetic variation down to the level of the single SNP, and to investigate how this variation is associated with human traits and diseases.

1.2.3 Family studies

Genetic studies based around families and patterns of inheritance have been used to investigate diseases that are common within families. Given a family pedigree, genetic marker information and disease information, the aim is to try to identify markers which cosegregate with the diseases phenotype and might therefore identify a region containing a causal genetic defect. A family-based linkage analysis may use a parametric model specifying the mode of inheritance, or a non-parametric model which assumes no inheritance model. A popular tool is the transmission disequilibrium test (TDT) (Spielman 1993). Heterozygote parents should pass on either allele to their offspring with equal probability. The TDT tests whether affected offspring carry certain alleles more frequently than would be expected given that they have heterozygote parents. The presence of alleles at high frequency in affected offspring could indicate that a causal variant near the tested allele is

inherited in affected individuals. Spielman et al applied this test to families with Type 1 Diabetes, classifying individuals into 3 allele groups based on an RFLP near the insulin gene. They discovered that one allele group was much more frequently inherited than the other two combined (p=0.004), giving evidence for linkage in this region. The TDT test is still useful today as it avoids problems which can occur in population studies, such as population structure. It is also used to identify regions for fine-mapping or sequencing to identify causal variants.

1.2.4 Candidate gene studies

Before it was possible to affordably and efficiently measure genetic variation across the genome, researchers often focused their investigations on a particular gene or set of 'candidate' genes they believed to be involved with a disease. Studies were then designed using this candidate gene list along with available information on SNPs in or near these genes. Individuals both with (cases) and without (controls) the disease were tested for polymorphisms at these SNPs. Allele frequencies in cases and controls were compared and tested, assuming that a large difference could indicate a SNP that was near a causal variant and therefore inherited with it. An advantage of a candidate gene study is that it is targeted at a small number of SNPs, and therefore fewer tests are carried out than in other approaches. This reduces the probability of false positive results due to chance.

This approach depends heavily on investigators' ability to predict which genes are associated with disease and may lead to missing associations that weren't predicted. Candidate gene studies have therefore not always successful at finding SNP-diseases associations in the absence of strong prior evidence for a particular gene or locus.

However if this evidence is available, this approach is still useful for fine-mapping loci identified by other means.

1.3 Genome-wide Association Studies

1.3.1 Genotypes and Phenotypes

A popular design for genetic studies is the genome-wide association study, or GWAS. Typically, genetic material is collected from the blood or saliva of study participants and 'genotyped' using a commercially prepared chip. Genotyping involves reading single bases of an individual's DNA sequence at many points throughout the genome. A base position in the genome where individuals vary is called a single-nucleotide polymorphism, or a SNP. The SNPs that are read by genotyping chips are generally picked to be on the chip because they are known to vary between individuals. Reading information at a locus where all individuals always carry the same base does not provide us with any information about differences between people. SNPs are also selected to try to capture as much common variation in the genome as possible. There are many SNPs which are not genotyped, but they may be 'tagged' by other SNPs in the region in high linkage disequilibrium; that is, with high r^2 values. A great deal of effort has gone into reading the sequence of many individuals to create a set of reference genomes for people of different ancestry (International HapMap Consortium, 2003, 1000 Genomes Project Consortium, 2010a). It has been shown that this genetic variability can help to explain why people display different physical characteristics, or phenotypes (Wellcome Trust Case Control Consortium, 2007). Phenotypes observed in individuals can be checked for association with genetic variability. This is the premise behind the genome-wide association study.

After a sample has been placed on a genotyping chip and processed, the data that is produced for analysis is a sequence of letters which represents an individual's alleles carried at many hundreds of thousands, or even millions of SNPs. The SNPs on a typical genotyping chip are usually bi-allelic; that is, there are only two alleles present at that locus in the population. Each individual at each locus caries two alleles; one from each chromosome. A person can carry two copies of one allele, two copies of the other allele, or one copy of each allele. A person carrying two copies of the same allele is a homozygote, while someone carrying one copy of each is a heterozygote. The alleles are often referred to by their relative frequency in the population; the allele at greater than 50% frequency is the major allele, while the other is the minor allele.

1.3.2 Quality Control

Genotyping is not a perfect science, and care must be taken in considering which SNPs to analyse. Poor genotyping and heterogeneous sampling can lead to many false association signals. There are many simple checks that can be carried out to ensure that the SNPs and samples that are analysed are of high quality (Weale, 2010, Anderson et al., 2010).

1.3.2.1 Sample Quality Control

Sample quality control aims to remove samples which were poorly genotyped, identify samples which may have been misidentified, identify ethnic outliers and identify samples from related individuals. Several summary statistics can be useful in finding such samples.

1.3.2.1.1 Sample missingness

Large amounts of missing data in a sample indicate a poorly genotyped sample. These samples are generally removed, as bias can be introduced if DNA quality varies with phenotype. This type of non-random missingness can lead to differential allele frequencies between cases and controls which may be interpreted as association signals.

1.3.2.1.2 Gender checks

Gender checks are a useful check of sample labelling. If the gender of individuals is available from the sample information, this can be compared to the genetically-determined gender. Large numbers of mismatches could indicate either incorrect sample labelling or an incorrect step in the data management process, and further investigation is required to ensure that the sample information and genotypes are correctly matched. Less prevalent mismatches could indicate problems within subsets of samples, for example mislabelling if samples from one centre. Gender checks can also be used to filter out rare instances of medical conditions such as Klinefelter's syndrome (XXY syndrome), in which a male carries an extra copy of the X chromosome. These samples with indeterminate gender are usually excluded.

Gender checks are usually performed by calculating proportions of heterozygosity for SNPs on the X-chromosome. Females with two copies of the X-chromosome will have SNP genotypes in Hardy-Weinberg equilibrium (HWE). A locus which is in HWE is showing the expected amount of heterozygosity given the allele frequencies for these SNPs and assuming random mating with respect to these loci. Males should have no heterozygosity on the X-chromosome as they carry only one copy; they should be hemizygotes (carriers of a single allele) for these SNPs. These

hemizygotes are typically coded as homozygotes in GWAS data sets and so will appear to be extremely divergent from HWE.

1.3.2.1.3 Relatedness and duplication

Duplicated and related samples will be more similar genetically to each other than would be expected in unrelated samples. This similarity could lead to differences in allele frequencies that could be confounded with phenotypes, leading to false association signals. Relatedness between individuals from the same population can be determined by working out the expected allele sharing between individuals given the allele frequencies of a SNP in the population. Averaging this value across all SNPs for a pair of individuals gives their 'identity by state' (IBS) value. IBS values can be converted to 'identity by descent' (IBD) values, which are estimates of the proportion of alleles they share due to recent common ancestors (Purcell et al., 2007). These pairwise IBD estimates will cluster around certain values for well-defined close relatives. Duplicates and monozygotic twins will have IBD of 1, while 1st-degree relatives (parent-child, full siblings) will have IBD of around 0.5. Second-degree relatives such as grandparent/grandchild, half-siblings, aunt/niece will have an IBD of about 0.25. Third-degree relatives such as cousins have an IBD of about 0.125. In a GWAS it is generally advisable to remove one of any pair with IBD somewhere between second and third degree relatives, and perhaps even lower if the resolving power to distinguish these cases from the general sample distribution of IBD is available. Generally the sample with the greater amount of missingness is removed. The remaining samples will then have a reasonable amount of allele-sharing as expected amongst unrelated individuals within the population being studied.

1.3.2.1.4 Ethnicity outliers

Alleles occur at different frequencies in different populations. Phenotypes can also be associated with different populations; for example cardiovascular disease has a prevalence of 15% (11%) in white males (females) and just 5% (5%) of Chinese males (females). Differences in ethnic makeup between cases and controls can therefore cause different allele frequencies at some SNPs, and these SNPs may be wrongly interpreted as associated with case/control status. This phenomenon is called population stratification. Genome-wide association studies are usually limited to a single population, and sample ethnicity can be verified genetically using Principal Components Analysis (PCA) (Price et al., 2006). More subtle population stratification can be adjusted for by including principal components that correlate with ethnicity in the genetic model that is used to test for association between phenotype and genotype.

1.3.2.1.5 Heterozygosity

An individual who is the offspring of unrelated parents from the same population should have genotypes in Hardy-Weinberg equilibrium. Higher than expected heterozygosity could be the result of sample contamination, while low heterozygosity could indicate that the parents are not from the same population, or could indicate an inbred individual. In any of these cases the unusual distribution of genotypes means that these samples should be removed from the analysis.

1.3.2.2 SNP Quality Control

SNP quality control aims to identify and/or remove SNPs that were poorly genotyped or that may be difficult to analyse. There is some flexibility in deciding to remove SNPs, as it is possible, and indeed recommended, to check quality control metrics for

any significantly associated SNPs after association analysis. SNPs showing strong associations but with poor QC metrics can either be ignored or given a low priority for replication.

1.3.2.2.1 SNP missingness

SNPs with high missingness should be removed from the analysis as this generally indicates poor quality genotyping. As with individual missingness, if DNA quality varies with the phenotype then non-random missingness can be mistaken for an association signal.

1.3.2.2.2 Minor allele frequency

SNPs with low minor allele frequencies (MAF) are often removed from the analysis or analysed separately using different methods. Low MAF SNPs are more sensitive to genotype calling errors as there is less information in rare genotypes for the clustering algorithm to use. Low MAF SNPs are also more sensitive to genotyping errors or non-random missingness, as small changes in rare groups can have large effects. Power is also reduced at low MAFs, so even when they are analysed it may be theoretically impossible, or very difficult, to get a strong enough association signal to overcome the required correction for multiple testing.

1.3.2.2.3 Hardy-Weinberg equilibrium

Extreme departures from Hardy-Weinberg equilibrium (HWE) can indicate poor genotyping. For example, poor separation in probe signal intensities between the three genotype groups can cause the calling algorithms to call only two groups instead of three. False association signals can result if this problem occurs preferentially in cases or controls, leading to a large difference in allele frequencies.

This is likely to happen more frequently if cases and controls are genotyped separately.

A very strong association signal in a case/control study can also lead to a departure from HWE in cases if the underlying risk model is non-multiplicative. It is therefore sensible not to be too stringent with the HWE threshold to prevent removing SNPs with a strong true signal or to apply HWE thresholding based on control genotype frequencies only.

1.3.3 Analysis of Association Studies

After genotype and phenotype data is collected, it now becomes a statistical challenge to judge if the two are associated. The type of phenotype being considered will determine the type of analysis carried out to test for association.

1.3.3.1 Association with Disease Status (Binary Trait)

1.3.3.1.1 Chi-square (allelic) test

A popular type of GWAS is the case/control design. People who are affected by a certain disease are designated as 'cases', and these are compared to controls who are either screened to be unaffected by the disease or an unscreened collection of people. We begin with a single SNP. In this scenario we can compare the presence of the two alleles in the populations of cases and controls. We can create a 2 x 2 contingency table and count the numbers in each cell as in Table 1.

Table 1. Example of a contingency table for applying a chi-square test to a single SNP

Allele	cases	controls
A	100	80
a	60	55

A simple chi-square test (with 1 degree of freedom) of association can be carried out on this table to test the independence of allele frequency and case/control status.

This is a test for a single locus, and it must be repeated for all loci in the study. Hundreds of thousands of tests can be carried out, and this dramatically increases the chances of seeing a positive results purely by chance. To deal with this problem a suitable correction for multiple testing must be applied, and this is discussed in section 1.3.3.3.

1.3.3.1.2 Trend Test

The Cochran-Armitage trend test is a modification of the chi-square test. Both the allelic and Cochran-Armitage tests are tests of a multiplicative allelic risk model, but the allelic test requires that the control genotypes be in HWE, while the Cochran-Armitage test does not so is generally preferred. For this test samples are divided into three genotype categories for each SNP as in Table 2.

Table 2. Example of a contingency table for applying a trend test to a single SNP

Allele	cases	controls
AA	100	80
Aa	60	55
aa	15	8

This is a very popular approach for a fast test of association, and has been used in several successful major studies (Wellcome Trust Case Control Consortium, 2007, Mells et al., 2011).

1.3.3.1.3 Logistic regression

Logistic regression is a more flexible regression modelling approach for binary outcomes, which allows additional covariates to be used. This allows, for example, for PC scores from a PCA to be added as covariates. The model is fit using maximum likelihood estimation to estimate coefficient values, as there is no analytical solution. Regression coefficients can then be used to determine the effect of each coefficient (e.g. a SNP) on the outcome (case/control status). This is usually done by a simple transformation of the coefficients to give odds ratios (the 'risk' of being a case given a particular genotype). An odds ratio is the ratio of the odds of an event occurring in one group (e.g. cases) to the odds of it occurring in another group (e.g. controls).

Equation 1 gives the form of the logistic regression model.

Equation 1. Logistic regression model with covariates

$$logit(p_i) = \ln\left(\frac{p_i}{(1-p_i)}\right) = \alpha + \beta_0 x_i + \beta_1 \mathrm{pc1}_i + \beta_2 \mathrm{sex}_i$$

where p_i is the probability of sample *i* being a case, α is a constant coefficient, βs are the regression coefficients, *x* is the genotype data, *pc1* is the first principal component, and *sex* is a dichotomous variable indicating the sex. Logistic regression is a widely used method in case/control designs, particularly when there are covariates to be added to the model (Strange et al., 2010, Evans et al., 2011).

1.3.3.2 Association with quantitative traits

Not all phenotypes of interest can be measured as a 2-category (dichotomous) outcome as in the case/control study design. Some phenotypes, such as height, weight, or the amount of a particular substance in the blood, are better measured on a continuous scale. If the measurements are normally distributed then we can analyse this type of data in a linear regression framework. The formula to describe this relationship is given by:

Equation 2. Simple linear regression model

$$y_i = \alpha + \beta x_i + \varepsilon$$

where y_i is the quantitative trait for individual i, α is a constant coefficient, x_i is the SNP allele count for individual i, β is the coefficient for the SNP, and ε is the error term.

1.3.3.2.1 Survival analysis with Cox Proportions Hazards modelling

Survival analysis is an extension of regression analysis to handle time-to-event data. In this scenario the time of the event of interest is recorded and this is used as the response. However it is possible that a subject can be observed over a time period

that does not include the event. Any such observations are considered to be 'censored' events. The event could have occurred before observation began (left-censored) or may not have happened by the time of the latest recorded observation (right-censored). In the case of right-censoring, a study could finish or a subject could be lost to follow-up before the event occurs. The last observation time is still recorded but the outcome is classified as right-censored. It is impossible to determine if the event would have happened the day after the recorded censoring time or many days, months or years in future, so it is not correct to view the censored time as an event time. However there is still value in knowing that the event has not happened in the time that has elapsed. In the data sets I have used for the work in this thesis only right-censoring has occurred, and no further reference will be made to any other form of censoring or truncation.

The basic elements used in survival modelling are the survival function and the hazard function. Let T be a continuous random variable with probability density function (p.d.f.) *f(t)* and cumulative density function (c.d.f.) *F(t)*. For a continuous random variable, the c.d.f. is defined as the integral of the p.d.f.

$$F(x) = Pr(X \leq x) \int_{-\infty}^{x} f(t)$$

The *survival function* can be thought of as the probability that the event of interest has *not* occurred by time *t*, which is the complement of the c.d.f..

$$S(t) = \Pr(T > t) = 1 - F(t) = \int_{x}^{\infty} f(t)$$

The *hazard function* is the probability of the event happening at the current time given that it hasn't happened yet. This is a conditional probability which can be defined for discrete and continuous time. For a discrete distribution, the hazard can

be calculated at each time point. For continuous time, the hazard at time t is calculated over a time interval (t, t+dt) as the size of the interval (t+dt) approaches 0. The probability of the event happening at time t is given by the pdf, while the probability of the event not happening until time t (the condition) is given by the survivor function. The hazard function λ(t) is therefore the ratio of these two values.

$$h(t) = \frac{f(t)}{S(t)}$$

If the hazard rate is constant over time, survival time could be modelled by an exponential distribution. However in real data it is likely that the hazard function may not be constant over time, and it might not be sensible to assume a underlying distribution. For example, a person's chance of dying might be higher at a very young age, then lower as the risk of childhood illnesses falls, then rise again as the risk of age-related disease rises. The Cox proportional hazards model (Cox, 1972) is a semi-parametric approach to modelling hazards. This model is considered semi-parametric because there is a baseline function which can take any form, but variables enter the model in a linear predictor. The baseline hazard in a Cox model also absorbs the constant term which is usually a part of a parametric regression model. The Cox proportional hazards model is:

Equation 3. Cox proportional hazards model

$$\log_e h_i(t) = \alpha(t) + \beta_n x_{in}$$

where $h_i(t)$ is the hazard function, $\alpha(t)$ is the baseline hazard, and $\beta_n x_{in}$ is the set of *n* coefficients (*ß*) for *n* variables (x) over *i* subjects. The baseline hazard is the regression constant, which is the value of the hazard function when all of the x's are 0 (e.g. $\alpha(t) = h_0(t)$). The remainder of the right-hand side ($\beta_n x_{in}$) is a linear predictor for which parameters β_n can be estimated.

Estimation of parameters is done by maximising what Cox called the partial likelihood function.

Equation 4. Cox's partial likelihood function

$$L_p(\beta) = \prod_{i=1}^{n} \left[\frac{\exp(x_i'\beta)}{\sum_{i' \epsilon R(t_i)} \exp(x_{i'}'\beta)} \right]^{c_i}$$

where $R(t_i)$ is the set of subjects at risk (*risk set*) at time t_i (event time for subject i, censored or not) and c_i is an indicator variable with c=0 for a censored observation and c=1 for an event. Censored times enter the likelihood function as part of the risk set but are not included in the likelihood as separate terms, as the entire term for censored times reduces to 1. The ratio $\frac{\exp(x_i'\beta)}{\sum_{i' \epsilon R(t_i)} \exp(x_{i'}'\beta)}$ is the hazard for subject i at that subject's event time, relative to all subjects at risk.

Maximum partial-likelihood estimates share many of the same properties as maximum-likelihood estimates, and can be used in likelihood ratio tests.

1.3.3.3 Correcting for multiple testing

Statistical tests such as described in section 1.3.3 are usually evaluated at a threshold of significance such as 5% or 1%. Under the null hypothesis of no association, p-values are uniformly distributed between 0 and 1. As a result, a threshold of 5% means that if the experiment were repeated 100 times, we would expect about 5 of the results to have a p-value of 0.05 or below. In a GWAS we will be carrying out not 100, but several hundred thousand tests. If there are 500 000 SNPs to be tested, we would expect 25 000 of these tests to give a 'significant' p-value of 0.05 or less even if there are no real associations at all, which is clearly not a desirable outcome. It is therefore important to set a very strict threshold, far smaller than 5%, to try to ensure that results are real.

A commonly used correction for multiple testing is the Bonferroni correction. This involves dividing the desired threshold for significance by the number of tests. In the case of a GWAS with 500 000 SNPs, measured at the 5% level of significance, this would mean that the Bonferroni-corrected threshold for significance is $0.05/500\,000 = 10^{-7}$.

1.3.4 Important results from Genome-Wide Association Studies

The first successful genome-wide association study was published in 2005 (Klein et al 2005). Researchers genotyped 95 people with age-related macular degeneration and 50 healthy controls at 105 098 SNPs after quality control measures were applied. This was a fortunate choice of phenotype, as most common variants are of small effect size, and would not be detectable with such a small sample size. This study found a SNP in the complement factor H (CFH) gene to be strongly associated with AMD (OR=4.6 (heterozygotes) and 7.4 (homozygotes), $p=4.1x10^{-8}$,). This SNP was intronic and therefore unlikely to be causal. Follow-up sequencing of CFH exons led to the discovery of the associated non-synonymous SNP rs1061170. This SNP is still being investigated with mixed findings and varying odds ratios; in particular it does not seem to be associated with AMD in non-Caucasians (Sofat et al 2012). It is still unclear whether this variant is causal or if it tags an undiscovered causal variant. Since allele frequencies and LD patterns vary between populations, one possibility is that that this SNP is in LD with the causal variant in European populations but not others, explaining the different findings in non-Caucasions.

In 2007 the Wellcome Trust Case Control Consortium published an important paper on a large-scale GWAS in 7 different diseases. The sample sizes for these studies were much larger, with approximately 2000 cases for each disease. It was also the first study to use shared controls, with a single set of approximately 3000 unscreened

population controls. Association signals were identified in five of the seven phenotypes studied, the most successful being Crohn's Disease with 5 associations passing a strict Bonferonni-corrected p-value threshold. This study showed that using large sample sizes, strict quality control, appropriate multiple testing correction and independent replication could result in success for GWAS.

Many GWAS have now been carried out, and the resulting associations have been collected and summarised into a catalogue of published GWAS by the National Human Genome Research Institute (Hindorff et al, 2009).

1.3.5 Epistasis

The definition of epistasis is complicated and depends somewhat on one's point of view (reviewed in (Cordell, 2002)). Biologists and statisticians will almost certainly give different definitions, with neither of these likely to be exactly what the people who first used the word in the respective fields intended. Here I will outline several definitions of the word and specify how I will use it in this thesis, as well as introduce various methods and software packages currently in use for their detection.

1.3.5.1 Measuring Epistasis

A definition of biological epistasis was first given by Bateson in 1909 (Bateson, 1909). He examined a family of rabbits that carried a gene for albinism. He describes two pairs of 'allomorphs' (alleles), one of which determines grey/black colour and the other determines colour/albinism. He noted that "The presence of one or other determiners G or B is only perceptible when it exists in combination with the colour-factor".

After explaining all of the possible combinations of alleles and their outcomes, he goes on to introduce new terms to describe this relationship:

"Pending a more precise knowledge of this relationship it will be enough to regard those factors which prevent others from manifesting their effects as higher, and the concealed factors as lower. In accordance with this suggestion the terms epistatic and hypostatic may conveniently be introduced."

The term epistasis continues to be used today, although the term hypostatic is no longer in common use. Bateson's use of the term epistatic to refer only to the effect of the 'higher', or masking, factor is rarely used now, and instead the term epistasis usually refers to the general situation in which one gene modifies the expression of another gene.

The word epistasis is now also commonly used in quantitative genetics, and was described in this sense by Ronald Fisher (Fisher, 1918). He described epistasis:

We may say that the somatic effects of identical genetic changes are not additive, and for this reason the genetic similarity of relations is partly obscured in the statistical aggregate. A similar deviation from the addition of superimposed effects may occur between different Mendelian factors. We may use the term Epistacy to describe such deviation, which although potentially more complicated, has similar statistical effects to dominance.

This is a rather long-winded way of saying that epistasis is the deviation of effects from the additive linear model. In section 1.3.3 the analysis of genetic data using a linear model is described. This model includes only 'main' effect terms; that is, the effects of each term in the model act independently. The direction and size of effect due to one term in the model do not depend on the value of any other terms in the

model. When this is not the case, including non-linear terms such as interactions in the model may better explain the relationship between genotypes and phenotypes. In the simplest two-locus case, we can look at counts of cases and controls across the two loci and visualise a non-linear relationship. Figure 1 shows the effect of carrying different alleles at two loci, A and B. The effect of the alleles carried at locus B varies depending on the alleles carried at locus A. People who have the AA genotype at locus A have no increased risk regardless of the genotype at B. However if a person carries one copy of the "a" allele then risk increases depending on the number of "b" alleles carried at locus B. This effect is even greater in those carrying two copies of the "a" allele.

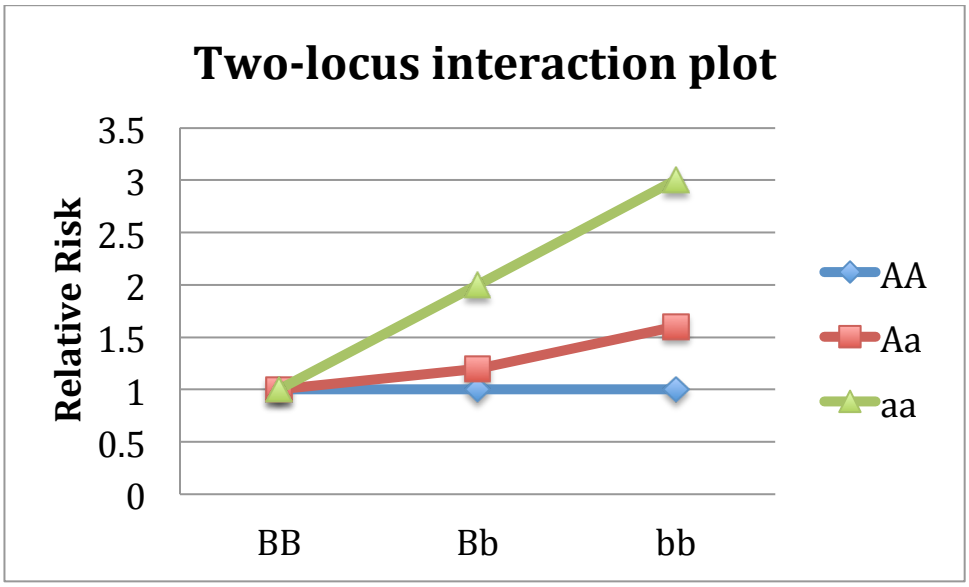

Figure 1. Two-locus interaction effect

Modelling this mathematically would require a non-linear term to describe the interaction between two effects, such as in Equation 5.

Equation 5. The linear model with two SNPs and an interaction term

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \varepsilon$$

where y_i is the response for individual i, β_1 and β_2 are the coefficients for the two variables x_1 and x_2, β_3 is the coefficient for the interaction between x_1 and x_2, and ε is the error term.

For the remainder of this thesis, the use of the word 'epistasis' will be restricted to the statistical definition above unless otherwise indicated. Furthermore, the terms "genetic interaction" and "gene-gene interaction" will be used synonymously with epistasis.

1.3.5.2 Interactions in Genome-Wide Association Studies

The hypothesis that groups of SNPs act together to affect a phenotype is a plausible one. It sounds reasonable biologically that a disease or a physical trait could be determined by a group of polymorphisms rather than just one.

There has been interest in detecting epistatic effects in GWAS data. These are widely presumed to exist (Zuk et al., 2012), and finding these would constitute an important source of extra information from our GWAS studies. There are many reasons why finding these effects in GWAS data is challenging, (1) the search space grows exponentially as we increase the number of SNPs, (2) power is reduced due to the increase in degrees of freedom arising in a model with interaction terms (particularly if the full genotypic model is used) and (3) poor tagging of causal SNPs has a greater effect on the power to detect higher-order interactions than main effects (Reimherr and Nicolae, 2011).

There have been few examples of statistical interactions between SNPs in a GWAS that have been successfully replicated in an independent cohort. Recent studies into psoriasis and ankylosing spondylitis, two immunological diseases, both discovered interactions between a SNP in ERAP1 and a SNP in the major histo-compatibility

(MHC) region on chromosome 6. The study in psoriasis (Strange et al., 2010) discovered an interaction between SNPs in ERAP-1 and HLA-C genes, and this was confirmed in their replication cohort. The study in ankylosing spondylitis (Evans et al., 2011) found an interaction between ERAP-1 and HLA-B, and this also replicated. In each case these interactions were found by taking all SNPs which showed genome-wide single-SNP association signals and testing for interactions between all possible pairs. The replication of each of these findings within their studies, and the similar findings in two immunological diseases, make these the most convincing interactions found to date in a GWAS.

1.3.5.3 Current Methods and Software for Interaction Testing

It is now computationally feasible to analyse all pairwise interactions in a GWAS (Marchini et al., 2005). However this approach requires significant computing power and is impractical if not impossible for higher-order interaction effects. Several more efficient methods have been suggested for searching for interactions in GWAS data (reviewed by Motsinger-Reif et al., 2008, Cordell, 2009). Many require some form of pre-screening to limit the number of interactions that need to be analysed, which usually involves the selection of SNPs based on single-SNP association p-values. This makes interaction testing quicker and easier but may miss interactions in which one or both SNPs lack a strong marginal effect. Here I will describe some of the popular methods and software packages and highlight their strengths and weaknesses. Chapter 2 of this thesis will describe a comparison of a selection of these methods for detecting epistasis in GWAS, with a focus on finding a software package that can efficiently analyse whole-genome data without a pre-screening step.

The software package PLINK (Purcell et al, 2007) has two options to test for interactions. The first fits a full logistic or linear model including an interaction term and tests if the coefficient of the interaction is equal to zero. The second is a faster, approximate test for case/control designs only. First the data is divided into cases and controls and odds ratios are calculated for a pair of loci in each group. Next a z-score is calculated for the difference between the odds ratios in cases and controls. A significant difference means that cases and controls carry different combinations of alleles across the two loci, and this may indicate epistasis.

Tree-based Epistasis Association Mapping (TEAM) uses a "minimum spanning tree structure", which involves building a contingency table for two SNPs. The table is pared down to utilise only those cells with different genotypes between the two SNPs. Removing the unnecessary samples reduces computation time and speeds up the analysis, however this method still doesn't scale well and can't be run on genome-wide data.

Multifactor Dimensionality Reduction (MDR) (Ritchie et al 2001) takes high-dimensional SNP data and reduces it to a single dimension by allocating each multi-locus genotype to high-risk or low-risk categories. The single-dimensional model is then used to predict disease status. Although popular, MDR does not scale up well to large numbers of SNPs such as from a GWAS and would therefore require a pre-screening step to run on GWAS data.

Bayesian Epistasis Association Mapping (BEAM) (Zhang et al, 2007) implements an MCMC simulation algorithm to estimate posterior probabilities for each SNP's membership in three groups - SNPs with no effect, SNPs with main effects only and SNPs which interact. As with many Bayesian simulation methods, BEAM is slow and it does not scale up to the number of SNPs measured in a GWAS.

Various penalized regression models have been applied to GWAS data to attempt to build stable models with large numbers of predictors using relatively small sample sizes. Although these methods can in theory fit parameters to high-dimensional data, they do not scale up to the number of parameters that would need to be modelled for a full pairwise interaction analysis using GWAS data.

INTERSNP is a software package which implements several possible screening steps to reduce the number of SNPs for analysis. This is followed by a choice of interaction analysis tools including parametric and non-parametric methods. Although this software can handle genome-wide data as input, it does so by first screening out some of the data before applying the interaction tests.

Since carrying out the work described in this chapter, further methods and software packages have been released that can be applied to GWAS data without a screening step. EpiBlaster (Kam-Tong et al 2011) uses a two-stage method. The first stage screens using a fast but lenient method, allowing strong interactions to affect the ranking of SNPs. Stage two specifically models interaction effects for SNPs passing stage one. Epiblaster is also implemented for use on graphical processing units

(GPUs) which can perform high-speed parallel processing on a single processor. BiForce (Gyenesai et al 2012) is a recently published software package which uses efficient storage structures, fast operations and parallelization to accelerate pairwise analysis. SIXPAC (Prabhu et al, 2012) is one of the fastest software packages for detecting epistasis in GWAS data, using ROC curve analysis as a first-stage screen which allows for interaction effects. EpiBlaster, BiForce and SIXPAC were not available at the time of carrying out the work in Chapter 2, but may be sensible current choices for detecting epistatic effects in GWAS data.

Random Jungle (Schwartz et al) is a fast implementation of the Random Forest algorithm (Breiman et al). Multiple binary trees are created by splitting data recursively. At each split point the data is divided by genotypes to produce new groups which maximise cases in one group and controls in the other. The stochastic element is introduced in two ways; limiting the samples available to build each tree and having only a subset of SNPs available at each split point. Random Jungle is fast and was built specifically to be used on GWAS data.

SNPHarvester is a heuristic hill-climbing algorithm which is guaranteed to find local maxima, but will not necessarily find the best global solution. Random SNP groups are selected to start the hill-climb, and new SNPs are tested one at a time for improved scores for association of the SNP group with the phenotype. This process is repeated many times, and SNP groups whose scores exceed a user-defined threshold are saved for follow-up. Longer run-times will search more thoroughly and increase the probability of finding the global maxima or significant local maxima.

Sparse Partitioning, in its original format, is a Bayesian method for identifying sets of predictors associated with a response. It also explores all possible 'partitions' of those predictors, where a partition is a group of predictors jointly affecting the response. In this way, single-SNP and interactions of an order up to a user-defined limit are explored at the same time and without assumptions about genetic architecture, making this an attractive option for exploring GWAS data. As with many Bayesian solutions it does not scale up to whole-genome data, but there is an alternate implementation using a deterministic approach which makes analysing GWAS data computationally feasible.

In Chapter 3 I describe work on comparing three methods for testing interactions, using a real data set with a replicated interaction effect as a positive control. I selected methods primarily based on their ability to process genome-wide data without a pre-screening step, to allow for the possibility of interaction effects in the absence of marginal effects. Ideally my optimal method should be able to efficiently search as much of the interaction search-space as possible, with reasonable power to detect interactions in typical genome-wide association studies.

1.3.6 Heritability

Variability in a trait between individuals is attributable to genetic and environmental effects. The heritability of a trait is defined as the proportion of total variance in the trait explained by genetics. Heritability of traits or medical conditions is usually calculated by analysing sets of twins. Twins may be monozygotic (sharing 100% of their DNA) or dizygotic (sharing 50% of their DNA). Because twins are likely to

share a large part of their environment, we can assume that differences in prevalence between mono- and dizygotic twins of affected individuals are due solely to differences in genetic makeup. This allows us to estimate heritability as a percentage. For example, a disease that is 30% heritable means that 30% of a person's risk of developing a disease is attributed to their genetic makeup, while the remainder is due to environmental factors. Examples of possible environmental factors contributing to disease risk are diet, smoking, urban vs rural environment and exposure to toxins.

Genome-wide association studies to date have successfully discovered many loci associated with diseases in case/control studies. However the amount of variability explained due to these associated variants is usually not anywhere near the heritability as estimated in twin studies (Maher, 2008). It does seem likely that a complex organism such as a human being will have a complex genetic architecture, with many genes working together to produce measurable phenotypes. Interactions between genetic variants have therefore been suggested as one possible contributor to missing heritability in complex genetic traits and diseases.

1.4 Kidneys, Renal Failure and Transplantation

Our kidneys are a key part of the urinary system, and their functions are essential to our health. The kidneys are responsible for many functions such as filtering the blood by removing certain toxins and excreting them as waste, maintaining blood pressure through balancing salt and water in the blood, reabsorbing water and some nutrients and producing some hormones and enzymes.

1.4.1 Kidney Failure and Dialysis

Around 1 in 10 people have some degree of chronic kidney disease (CKD) (British Kidney Patient Association, 2012b). CKD is a descriptive term for kidneys with long-term poor function. CKD is not a disease itself; it is usually a response to an underlying condition, and can sometimes be reversed depending on the cause. If renal function deteriorates beyond a certain point then the person has 'end-stage renal failure' (ESRF), and renal replacement therapy is indicated. Approximately 10% of people with CKD progress to renal failure. Some of the common underlying causes of renal failure are diabetes, high blood pressure and polycystic kidney disease. People with the above conditions are not all on the transplant list; treatment can maintain or improve renal function in some cases.

In some situations a kidney can suffer acute damage (British Kidney Patient Association, 2012a). Acute renal failure is often temporary. It can be caused by loss of blood supply to the kidney (for example, due to low blood volume after an injury), toxins or obstruction of the urinary tract. If the underlying cause can be treated successfully then kidney function may return. However permanent kidney damage is a possibility, in which case the individual can have end-stage renal failure.

Once a person is in ESRF the patient must receive some form of renal replacement therapy, or the build up of toxins and the imbalance of fluids and salts in the body can be fatal. Patients may undergo dialysis, which involves mechanically and chemically filtering the blood.

Traditional haemodialysis involves visiting the hospital up to three times per week and having a large volume of blood removed, filtered and returned to the body (British Kidney Patient Association, 2012c). It usually involves the insertion of a

catheter or construction of a fistula, both of which can be painful and inconvenient. Home dialysis is sometimes possible but involves a significant amount of training and a person to help with the home dialysis.

Peritoneal dialysis (British Kidney Patient Association, 2012d) can be done at night (automated peritoneal dialysis, or APD) or during the day (continuous ambulatory peritoneal dialysis, or CAPD). The lining of the abdomen, or the peritoneum, acts in place of the kidney. Fluid is introduced into the peritoneal cavity and sits for several hours. Waste products pass from the peritoneum into the fluid, then the fluid, with the waste, is drained away. This process can be carried by the patient out at home, or in any clean environment with the right equipment. This option may allow the patient more freedom than traditional haemodialysis.

The aim of dialysis is to keep the patient alive, but there are side effects and risks to consider. Patients may suffer from side effects such as tiredness, nausea, loss of appetite and depression. A dialysis patient's fluid intake and diet needs to be tightly controlled. There is a risk of infections at the catheter insertion site or fistula (in haemodialysis), or the peritoneum (in peritoneal dialysis).

1.4.2 Kidney transplantation and the UK Transplant Waiting List

A patient with end-stage renal failure will usually be put on the transplant waiting list. If a transplant is possible, the outcome is generally far better than long-term dialysis. While dialysis can keep a patient alive, a transplant can restore renal function and with it a person can return to full fitness. A transplant itself is not without risk but the long-term prognosis is much better than staying on dialysis. A large American study of patients on long-term dialysis for end-stage renal failure

found that the long-term mortality risk for transplant recipients was 68% lower than that of patients on the waiting list (Wolfe et al., 1999).

Ninety per cent of people on the UK's transplant waiting list are waiting for a kidney. The number of people currently waiting for a kidney is about 7000, yet only 2 500 transplants are carried out each year (Johnson et al., 2010). Transplanted grafts do not necessarily last for the lifetime of the patient, and when the kidney fails the recipient will have to go back on the transplant waiting list and back on dialysis while waiting for a new kidney. A recent large study of transplants involving deceased donors (Wolfe et al., 1999) found that death-censored graft survival at 5 years was 85.1% for cardiac-death donors and 83.2% for brain-death donors. Another recent study of deceased-donor transplants (Watson et al., 2012) found that 5-year graft survival was 76.7-83.1% after stratification by a number of risk factors. The most recent report from NHS Blood and Transplant (the organisation which maintains a database on all UK transplants) found five-year graft survival following cadaveric donor transplantation to be 79-83% for brain death donors and 73-85% for cardiac death donors, while ten-year survival was 67% for both groups.

These patients whose transplants fail will go back on the transplant waiting list, creating more demand for a very limited resource.

1.4.3 Factors Affecting Long-Term Transplant Outcome

Before discussing the impact of various clinical factors on transplantation, there are several key terms relating to transplantation that need to be defined:

A *graft* is a term for living tissue that has been transplanted. An *allograft* is any transplanted tissue that is from a source that is not genetically identical (e.g. not from self or identical twin).

A *heart-beating donor* (HBD) has suffered brain death but is kept alive on life support until organs are removed for transplantation. Circulation is maintained until the time of organ harvesting. A *non-heart-beating donor* (NHBD) is not sustained on life support at the time of organ harvesting; they have already experienced cardiac death. They may be brought to hospital already deceased, they may have had cardio-pulmonary resuscitation on the way to or at the hospital, or they may undergo cardiac death in hospital after the withdrawal of life support, but before organs are transplanted.

Cold ischaemic time is the time between the removal of a kidney from a donor and the time blood flow is restored to the kidney after transplantation in the recipient.

HLA type - The current nomenclature for HLA typing describes the alleles a person carries in the HLA genes. The type for each protein is preceded by the locus name (e.g. "HLAC") and followed by an even number of digits which come in pairs. The type can be of variable 'depth', with the shallowest description containing 2 digits that describe a broad molecular type which groups many alleles together. As technology has advanced, more specific typing has become possible and further pairs of digits (up to 4 pairs in total) now identify HLA type sub-groups more specifically. Kidney donors and recipients in the past have been matched as closely as possible by HLA type as rejection of the organ was less likely. More recently, better immunosuppressive drugs have meant that poor HLA matching can still lead to a good transplant outcome, and this is becoming more common (Johnson et al., 2010).

Serum creatinine is used as a biomarker for kidney function. Creatinine is produced by muscles and is filtered out of the blood by the kidneys. High creatinine levels indicate poor kidney function. Creatinine levels that do not come down after transplantation or that rise sharply in a transplant recipient may indicate an episode

of *acute rejection*. Creatinine levels that rise slowly over time may indicate *chronic rejection*.

Acute rejection is caused by an immune system response to the transplanted organ. It is usually suspected when a recipient's serum creatinine rises sharply, usually within the first year of transplantation. It can usually be treated with a strong dose of immunosuppressants and kidney function may return to normal after an acute rejection episode, though it doesn't always.

Chronic rejection (also known as chronic allograft nephropathy) is a slow process which results in damage to the kidney and may eventually lead the graft to fail. A change in the long-term immunosuppressive regime may slow or stop graft deterioration but this is not always possible.

1.4.3.1 Clinical Factors

An understanding of factors affecting the graft survival time for a kidney transplant may be a useful first step in relieving the pressure on the transplant list. Matching donors and recipients as well as possible may help prolong the life of a graft, which will mean less people returning to the waiting list. Much work has been carried out to identify clinical factors in both donors and recipients that may affect the long-term prognosis of the transplanted graft. A recent UK study into the effect of heart-beating vs non-heart-beating donors (Summers et al., 2010) in kidney transplants found no differences in survival between the two groups. Within the non-heart-beating group, they also showed that increasing age of donor and recipient; repeat transplantation and longer cold ischaemic time were associated with worse graft survival, while poor HLA matching had a negative but not significant effect. A similar study involving deceased heart-beating donors (Johnson et al., 2010) found

older donor and recipient age, waiting time to transplant over 2 years diabetes and earlier year of transplant, HLA mismatching and cold ischaemic time were associated with poorer graft survival over 5 years.

1.4.3.2 Genetic Factors

Few genetic studies have been carried out in renal transplantation in relation to graft survival time. Genetic studies in transplantation to date have tended to focus on prediction of rejection using biomarkers, or analysis of gene expression data rather than single-nucleotide polymorphisms. Some genetic studies have also been carried out in renal failure in the absence of transplantation. No genome-wide association studies involving donors and recipients have been published in renal transplantation to date. Those genetic studies of polymorphisms in transplantation that have been carried out have been candidate gene studies, which have met with mixed success. Candidate gene studies have many challenges that make finding convincing results difficult. According to Ioannidis et al (Ioannidis, 2005):

"*...a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser pre-selection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance*."

It could be argued that all of these are true to some extent in many candidate gene studies. Small studies and small effect sizes are particularly prevalent in candidate gene studies, and both of these problems mean that published findings are more

likely to be false positives. For any particular outcome (such as a phenotype in a genetic study), the theory is this:

1. The pool of non-associated variants is very large, but a small proportion of these will wrongly be found to be associated by chance (false positives). This proportion is not affected by the power of the study to detect true associations.
2. The pool of true associations is probably very small, and the proportion of these that are found depends on the power to detect them. Power in candidate gene studies is often hampered by low sample size and small effect sizes, meaning few true associations will be found.
3. As the power to detect true effects decreases and the number of false positives remains constant, the proportion of false positives will be higher in the total pool of significant results, and these are the results that are likely to be published.

In short, it is likely that at least some low-powered studies that report significant findings are reporting false positives. Correcting for multiple testing (section 1.3.3.3) is essential and will certainly improve, but not eliminate, this situation.

In spite of the drawbacks outlined above, I will review the findings from genetic studies in renal transplantation bearing in mind that some results may be false positives, This may explain why many of these results are only published once, or subsequent studies have different findings.

Lee et al (Lee et al, 2012) looked at the APOL1 gene, which had previously been associated with kidney disease in African Americans (Genovese et al., 2010, Tzur et al., 2010), and did not find any association with graft survival time after

transplantation, although the study was limited to 119 African American recipients. The variants found in the African American studies are not present in European populations.

Wahrmann et al (Wahrmann et al., 2011) determined gene copy number variation of C4 (a component of the complement system) for donors and recipients in kidney transplants, hypothesizing that complement could play a role in graft survival. However they did not find any associations of C4 copy number with graft survival or any other unfavourable outcomes.

Winkelmayer et al (Winkelmayer et al., 2004) looked at three SNPs in the methylenetetrahydrofolate reductase (MTHFR) gene and found that certain genotype combinations across these loci were more frequent than expected by chance in recipients of kidney transplants who had stable graft function. These genotypes were not tested for differences in survival time.

Several candidate gene studies have been carried out on polymorphisms in genes coding for cytokines.

In one study of 199 transplant recipients, a variant in the IL-6 gene was associated with 5-year survival (Kocierz et al., 2011). Another study of the same variant in 335 recipients showed no association (Sánchez-Velasco et al., 2010).

Müller-Steinhardt et al (Müller-Steinhardt et al., 2004) genotyped recipients at 3 SNPs (including the previously-mentioned IL-6 variant) and analysed the resulting 5 haplotypes that existed in their samples across these loci. They discovered a common haplotype (GGG/GGG, frequency 28.5%) with an increased 3-year survival rate (95.6% vs 67.3 for all other haplotypes, $p = 0.0003$ for the log-rank test), which remained strong ($p=0.006$) after accounting for non-genetic covariates (presence of

antibodies, HLA-matching, year of transplantation, acute rejection, age and sex of donors and recipients). A follow-up study (Müller-Steinhardt et al., 2007) by the same group found that this haplotype was associated with lower IL-6 secretion in healthy individuals. A further study (Schulte et al., 2011) in potential kidney transplant recipients who were haemodialysed showed no impairment of IL-6 production in GGG/GGG individuals, though IL-6 production in the haemodialysed individuals was already significantly lower than in healthy controls.

In summary, candidate gene studies have been carried out but have shown mixed, mainly unconvincing results in searching for polymorphisms associated with long-term graft survival, though there may be some evidence of association near the IL-6 gene. To date there have been no published genome-wide association studies of donors and recipients in kidney transplantation.

1.4.4 Kidney Transplantation - Two Genomes in One

Transplantation creates a unique situation in which tissues with two different genomes exist in one body. The only other situation where this is likely to occur is between a mother and foetus during pregnancy. The presence of 'foreign' tissue in the recipient is likely to be a contributing factor in graft rejection, and so it is particularly interesting to consider the relationship between the two genomes. Variants in the recipient, the donor graft, or interactions between the two could conceivably affect the survival time of the graft, and searching for these interactions is one of the aims of this thesis.

1.5 Project Aims and Thesis Outline

The main aim of my PhD project is to search for statistical interactions in genome-wide association studies, particularly with reference to a study in renal transplantation.

To explore the nature of these interactions, I began my PhD with a project to search for interactions in GWAS data between SNPs in genes that coded for previously published protein-protein interactions. The hypothesis was that, for a given phenotype, these SNPs may be enriched for interaction signals compared to random SNP pairs throughout the genome. To explore this idea I wrote a permutation-based software package called PERSI (PErmutations analysis for Statistical Interactions), which I then applied to a WTCCC study in Crohn's disease.

I carried out a methods comparison in order to assess existing methods designed to search for statistical interactions in case/control GWAS studies. Many methods exist, and many have been evaluated and compared using simulated data. The recent discovery of the statistical interaction in a psoriasis GWAS (Strange et al., 2010) provided me with a useful positive control to allow me to test promising methods on a real data set.

Throughout my PhD I was involved in two studies in renal transplantation; a GWAS carried out at King's College London (KCL) and the WTCCC3 study in renal transplant dysfunction (RTD). The WTCCC3 RTD project was an expansion on the initial KCL study, and includes the data from this study. The main outcome of interest for these studies is the survival time of the transplanted kidney, and I was particularly interested in finding interactions that may influence this. This is a

particularly interesting problem due to each transplant involving two genomes - the donor and recipient.

1.5.1 Outline of the Remainder of this Thesis

Chapter 2. Searching for genetic interactions in psoriasis.

I have carried out a methods comparison of Random Jungle, SNPHarvester & Sparse Partitioning, using the HyperLasso method as a post-processing step for SNPHarvester and Random Jungle. I have applied these methods to real data from a psoriasis study (Strange et al 2011). This chapter will present the theory, software descriptions and findings of this study.

Chapter 3. Searching for Evidence of Genetic Interactions in Protein-Protein Interactions.

This chapter will present the theory, software implementation and findings of a study looking for an enrichment of low p-values between SNPs in genes coding for proteins involved in reported protein-protein interactions.

Chapter 4. Genome-wide association study in renal transplantation.

This chapter will outline my contribution to this study, including work on quality control and several case/control analyses

Chapter 5. Wellcome Trust Case Control Consortium (phase 3) study in renal transplant dysfunction.

In this chapter I describe my work on this large collaborative study, including quality control, case-control analyses of several secondary phenotypes, and survival analysis. I apply the best method from the psoriasis methods comparison study to this data to search for interactions. I will also present my work on the development

of the selected method from the methods comparison, expanding it to take survival data by implementing a log-rank function for time-to-event data.

Chapter 6. Discussion and Conclusions

2 Searching for Genetic Interactions in Psoriasis - a Methods Comparison

2.1 Introduction

Psoriasis is a complex immunological disease that has been the subject of a number of genetic studies. Several loci were found to be associated with psoriasis from linkage studies (reviewed in Elder et al 1994, Trembath et al 1997, Nair et al 1997). Genome-wide association studies narrowed the search and identified new loci (Cargill et al 2007, Nair et al 2009, Zhang et al 2009). In 2010 the Genetic Analysis of Psoriasis (GAP) consortium and the Wellcome Trust Case Control Consortium (phase 2) published the results of a large GWAS (Strange et al 2010), which replicated 9 previously-identified loci. They also discovered 8 new loci and replicated these in an independent cohort. All pair-wise interactions between these 17 loci were tested using logistic regression modelling, and an interaction between SNPs rs10484554 (HLA-C) and rs27524 (ERAP1) showed a significant interaction effect ($p=2.45*10^{-5}$, replication $p=0.027$, combined $p=6.95*10^{-6}$).

2.2 WTCCC2 Data

As part of the Wellcome Trust Case Control Consortium phase 2 (WTCCC2) psoriasis study, 2178 psoriasis cases were genotyped on the Illumina Human660W-Quad chip and compared to 5175 common controls genotyped on the Illumina custom Human1.2M-Duo platform (Strange et al., 2010). Analysis was performed on the overlapping set of 535475 SNPs after quality control filters were applied. An initial single-snp analysis replicated 9 previously-identified loci and 8 new loci and replicated these in an independent cohort. These association signals were found

using Cochran-Armitage trend tests (see glossary) which test the association of the number of alleles carried (0, 1 or 2) with case/control status, taking into account the direction of effect. For example a strong signal may be detected for a SNP that has few cases with no risk alleles, more cases with one copy of the risk allele and the most cases with two copies of the risk allele.

Strange et al used a single PC axis to control for population stratification, and I used the same PC axis (PC1) in all of my analyses. Strange et al. checked for interactions between all of the SNPs whose single-SNP association p-values were smaller than the Bonferroni-corrected threshold for genome-wide significance. An interaction was found between SNPs in the HLA-C gene in the Major Histo-compatibility Complex region (MHC region) on chromosome 6 and the ERAP1 gene on chromosome 5. The interaction was replicated in an independent set of cases and controls. This confirmed interaction will act as a positive control in my study.

2.3 Methods

In the Introduction to this thesis I described several methods that have been developed to test for interaction effects between SNPs in GWAS data. The aim of this piece of work was to evaluate methods that can search the entire set of GWAS data without resorting to screening the data to reduce the search space. This lead me to select three software packages which could run on full GWAS data sets with the computing hardware available to me at the time of carrying out the work. Since this work was carried out, several software packages have been released which use the power of graphics processing units (GPUs) to speed up analysis. Programmable GPUs are now more widely available, and their use may make some of the slower methods and programs applicable to GWAS data.

2.3.1 Random Jungle

Random Jungle (Schwarz et al., 2010) is a fast implementation of the Random Forest algorithm (Breiman et al), which is a method based on the binary decision tree. Random Forests is an ensemble method, meaning many trees are built on *bootstrapped* samples, or samples of the same size taken with replacement.

This creates many decorrelated trees, and averaging across these trees provides better prediction than a single tree. Bootstrap sampling leaves approximately 1/3 of the samples out of each tree ("out-of-bag" or OOB samples), and these OOB samples can be used to calculate prediction accuracy.

Each tree in Random Forests is formed by recursively splitting the data in two to decrease an impurity measure in each of the two new "daughter" nodes. The change in impurity measure is calculated by:

Equation 6. Change in node impurity

$$\Delta i = i_{parent} - (p_{left} * i_{left} + p_{right} * i_{right})$$

where p_{left} and p_{right} are the proportion of samples in the left and right daughter nodes, respectively, and i_{node} is the Gini Index for that node. The Gini index is a commonly used measure of homogeneity (see glossary for formula).

Since our SNPs have three classes, there are different ways in which a SNP can be used to split data into two groups. The algorithm chooses the best binary split after either grouping heterozygotes with the major homozygote class or with the minor homozygote class. Resulting daughter nodes are then further split in the same way to increase purity. Nodes are split this way down through the tree until terminal nodes are 100% pure. Once classification trees are grown they can be 'pruned' back to the

point at which the splits significantly improve classification, but this is not done in Random Forests; trees are left fully grown, with pure terminal nodes.

A randomly-determined subset of SNPs is available at each node for selection as the split variable. The size *m* of the available subset is determined empirically to reduce classification error in the out-of-bag samples, as suggested by Schwartz et al. The size is held constant throughout the procedure, while a new subset is chosen at every node.

I calculated variable importance measures (VIMs) using the Meng score (Meng et al., 2009). First, OOB samples for every tree are classified using their real genotypes. The genotypes for a single SNP are then permuted and the samples reclassified. The Meng score for a SNP is based on the classification of the OOB samples using the real genotypes and the permuted data for that SNP, as shown in Equation 7.

Equation 7. Meng score for Variable Importance Measure (VIM)

$$I'_T(A) = \frac{2}{T_v}\sum_{j=1}^{T_v}\frac{1}{N_j}\sum_{i=1}^{N}[1\left(V_j(X_i) = y_i\right) - 1\left(V_j\left(X_i^{(A,j)}\right) = y_i\right)]t_{ij}$$

$I_T'(A)$ - importance measure for SNP A over all T trees, as given by Meng

T_v - number of trees containing SNP A

N - number of samples

$V_j(X_i))$ - classification of SNP A by tree j using genotype data X from all i individuals

y_i - true class of SNP A

$V_j(X_i)^{(A,j)})$ - classification of SNP A by tree j using SNP data from all i individuals, with SNP A genotypes permuted

The first term in parentheses will be one if classification using all genotypes is correct, and zero if it is incorrect. The second term reduces in the same way but using the permuted data for SNP A and real genotype data for all other SNPs. $I_T'(A)$ will reduce to approximately 1 if classification is perfect with SNP A and completely random when SNP A is permuted. This is perfect a Mendelian trait with 100% penetrance; it is unlikely that a genome-wide association study would be carried out on such a phenotype, so Meng scores are not likely to be as high as 1. $I_T'(A)$ will reduce to approximately 0 if real and permuted values for SNP A lead to the same classification; that is, SNP A is not associated with case/control status, so classification depends on all other SNPs. Most negative numbers will be small, being the result of slightly more correct classifications using permuted SNP A data than real SNP A data. Larger negative numbers would indicate systematically better classification when SNP A is permuted than when it is not, which is unlikely.

Previous studies (Nicodemus et al., 2007) have shown that VIMs are unstable and best estimates are given by averaging over as many trees as possible. I set Random Jungle to produce 500 forests of 500 trees each, and the median VIM was used across all trees to rank SNPs.

2.3.2 SNPHarvester

SNPHarvester (Yang et al., 2009) is a stochastic, hill-climbing algorithm designed to search efficiently through the space of all possible groups of SNPs. In this study I specified groups of size 2, thereby limiting ourselves to searching for pairwise interactions.

SNPHarvester does a 'first pass' at the data which removes any individual SNPs with large marginal effects based on a chi-square test of a full 2-d.f. genotypic model,

using a Bonferroni significance correction for all SNPs tested. These SNPs are removed from further analysis, on the assumption that they are easy to identify and can be investigated and tested for interactions without the need for sophisticated methods. The PathSeeker algorithm, which is the basis of SNPHarvester, then randomly selects a pair of SNPs and carries out a score test. For a case/control design this is a genotypic chi-square test (8 d.f.) of association with disease status. The algorithm then cycles through the rest of the data one SNP at a time, checking to see if the new SNP can replace one of the current SNPs to improve the score. If it does, the new SNP becomes part of the current group and the old SNP is removed. When all SNPs have been considered, if there have been any replacements then the cycle is repeated, again retaining any SNPs which improve the score. When the algorithm reaches the end of the data without replacing any SNPs then it has reached a local maximum. Any SNPs which exceed the score threshold (8-d.f. χ^2 values after Bonferroni significance correction for all possible SNP pairs) are considered significant and removed from the search space for future PathSeeker runs. The algorithm then begins again with a new randomly-selected pair of SNPs. PathSeeker runs until a pre-specified stopping rule is reached; I set SNPHarvester to stop when PathSeeker ran for 15 iterations without finding any new significant pairs. Figure 1 illustrates 4 runs of the PathSeeker algorithm (figure from Yang et al 2009).

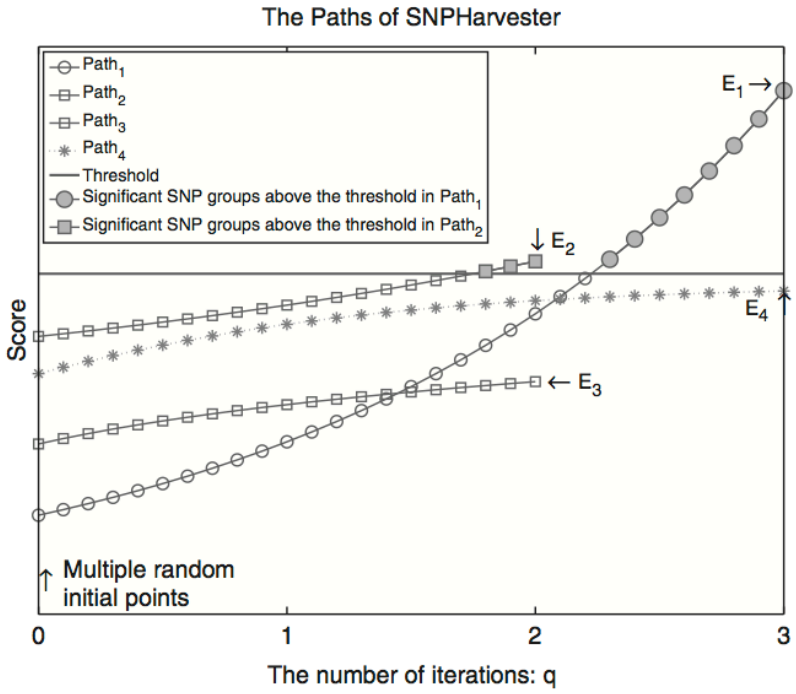

Figure 2. The PathSeeker algorithm

2.3.3 Post-processing with the HyperLasso

SNPHarvester and Random Jungle rank SNPs according to a measure of their influence allowing for interactions, but these methods do not specifically test the interaction effects. Therefore, the top 1000-ranked SNPs from these methods were put through 100 iterations of the HyperLasso (Hoggart et al., 2008), along with all of their 2-way interactions, in order to select the best terms for the model describing the association with the phenotype. I chose to take this number of SNPs and interactions mainly for practical purposes, as a reasonable number of terms to take forward to the HyperLasso. To select the Random Jungle SNPs, they were simply ranked by their variable importance measures and the top 1000 selected. For SNPHarvester, SNPs were ranked by p-value regardless of whether the p-value was generated in the initial single-SNP screening or the pairwise analysis. This allows for interaction effects in

the presence or absence of marginal effects. Each SNP in a pair was assigned the same p-value, so that both SNPs of any pairs would be selected together if they were ranked in the top 1000.

The HyperLasso algorithm is based on a penalised maximum-likelihood approach, and the penalty can be considered to be similar to a prior in a Bayesian framework. The goal of the HyperLasso is to maximise:

$$\log p(\beta|x, y) = L(\beta) - f(\beta)$$

where:

β - the set of regression coefficients (e.g. the SNP effects)

x - the set of genotypes for all individuals over all SNPs

y - the case/control status of all individuals

$L(\beta)$ - the likelihood of β

$f(\beta)$ - a function of β, applied as a penalty

I used a Normal-Exponential-Gamma prior and approximated the type-I error rate as suggested by Hoggart et al. The values of β are optimised using Newton's method. Notably, if a β changes sign (e.g. passes 0), it is very unlikely it will be able to move from zero, and terms with zero coefficients are removed from the final model. A non-zero coefficient is considered to be a significant term.

The HyperLasso software limits the genetic data to a count of minor alleles carried by an individual at a single SNP; that is, the data must be 0, 1 or 2. In our interaction model I have assumed that genetic effects are additive (on the log scale of the logistic regression model) and the interaction term is therefore the product of the number of minor alleles carried by an individual at two SNPs. This could be 0, 1, or 2 but also 4 in the case of a person who is homozygous for the minor allele at both

loci. I therefore amended the HyperLasso software to accept 4 as a valid data value. This allows the interaction data to be used, but the software will not force the main effects of an interaction term into the final model. In order to verify interaction terms, I therefore fit a logistic regression model for each interaction term selected by the HyperLasso including the main effects, then I dropped the interaction term and carried out a 1d.f. likelihood ratio test.

Over the iterations of the HyperLasso, different SNPs that are in high LD may be left in the model to account for the same signal. Comparing results across iterations may therefore miss strong signals because the signal is divided between these SNPs. To avoid this, SNPs were assigned to LD blocks based on an r^2 value of 0.5 or greater and these blocks were examined for overlap across iterations instead of the individual SNPs (Figure 3). I began at the first SNP on each chromosome and progressed along the chromosome, checking LD with the previous SNP. If the r^2 value was 0.5 or greater the SNP was added to the current group; otherwise a new group was created with the new SNP. Each group was then assigned an identifier in the form of chr1block1, meaning the first block on chromosome 1. Groups can have as few as a single SNP and there is no upper limit to the number of SNPs. These new chromosome/block terms were then used to amalgamate results across iterations and to check for overlap between the SNPHarvester and Random Jungle results.

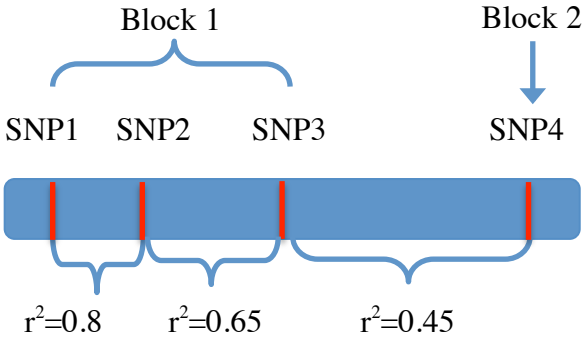

Figure 3. Assignment of SNPs to blocks

To account for population stratification in the HyperLasso, the first principal component was added to the model as a non-genetic covariate. This was also included in the follow-up logistic regression modelling.

2.3.4 Sparse Partitioning

Sparse Partitioning is a Bayesian regression method for identifying groups of associated SNPs. The method attempts to find groups of SNPs associated with a phenotype and define the best 'partition' of the predictors. For example a group of three predictors which are associated with a response could consist of three independent main effects, one pairwise interaction with a main effect, a three-way interaction, or, if using a SNP in multiple partitions is permitted, two interaction effects (e.g. SNP1 * SNP2 and SNP2 * SNP3). The user provides a prior parameter defining how many associated predictors there are likely to be in total, the possible depth of interactions (such as 2 for pairwise interactions only or 3 for 2-way and 3-way interactions) and if SNPs can be used in multiple partitions. The solution is assumed to be sparse, meaning only a few predictors are likely to be causal. Larger priors for the total number of SNPs and interaction depth will considerably slow down the computation of the solution.

The original Sparse Partitioning algorithm is a stochastic search, meaning that it randomly searches through the space of all possible partitions, moving predictors into and out of groups at each iteration; at the end it calculates posterior probabilities of association and pairwise probabilities of interactions by recording how often each predictor was in a non-null group and how often two predictors were in the same

non-null group. This solution does not scale up for whole-genome data; the computational burden is too great. Instead I used the modified deterministic version (Speed, 2010) which is somewhat similar to stepwise regression. For each of a sequence of moves, a SNP is either added, removed or swapped to the group of predictors associated, and the process stops when the score cannot be improved. This method allows the user to specify an exact number of associations. If the best solution does not contain this number of SNPs the process will restart, either adding or removing predictors until it selects the specified number.

For this study Sparse Partitioning was set to choose four predictors, and up to two of those could be two-way interactions, hence a total of 4 to 6 SNPs would be in the final group of associated SNPs. The number of predictors in the final model is chosen mainly for practical reasons, as larger models dramatically increase computational time.

The first principal component, as used by Strange et al, was given a prior probability of 1 to force it into the model. The top SNP in the MHC region was also given a prior of 1, as the effect is very large and not in doubt. Sparse Partitioning was sensitive to genotyping errors and was initially selecting only SNPs that proved to be badly called in either the cases or the controls, creating large differences in allele frequencies between them. To avoid this problem I set a more stringent threshold for Hardy-Weinberg equilibrium ($p<0.004$) and applied it to cases and controls, which removed the problem SNPs.

2.3.5 Replication

I attempted replication of all SNPs involved in pairs that were selected by all 100 iterations of the HyperLasso on the SNPHarvester results. When more than one SNP

pair contributed to an LD block I attempted to replicate every pair. Data from the Collaborative Association Study of Psoriasis (CASP) study was used for replication (Nair et al., 2009). This study genotyped samples on a Perlegen platform (see Nair et al. for details). SNPs for replication of the SNPHarvester interactions which were not genotyped were imputed with IMPUTE2 software (Howie et al., 2009), using data from the July 2011 release of the data from the1000 Genomes Project (1000 Genomes Project Consortium, 2010b).

I also attempted replication of a novel single-SNP association selected by Sparse Partitioning. This SNP had a perfect proxy SNP (r^2=1) in the CASP study and this was used for replication.

For replication of interactions, a logistic regression model was fit to the imputed CASP data with the interaction terms of interest and their main effects. I then fit a second model without the interaction term and carried out a 1 d.f. likelihood ratio test to test if the interaction term accounts for any extra variance beyond that explained by the main effects. The Bonferroni corrected level of significance required accounted for all of the terms in all of the models in the replication stage. A logistic regression model was also fit using the proxy SNP from the CASP data to replicate the novel SNP found by Sparse Partitioning. To test the significance of the SNP I dropped it from the model and carried out a likelihood ratio test.

2.4 Results

2.4.1 SNPHarvester

SNPHarvester results can be divided into single-SNP associations removed before pairwise analysis and pairs selected for their combined association with psoriasis case/control status. Seven hundred and ninety-five SNPs were removed in the first

stage with strong single-SNP p-values, and 3593 SNP pairs were selected by SNPHarvester. In total from both single-SNPs and pairs there were a total of 1549 unique SNPs selected by SNPHarvester. Table 3 gives the p-values for SNPs that were discovered or replicated in the WTCCC study alongside the SNPHarvester rankings. A large proportion of the SNPs selected by SNPHarvester were in the MHC region, so the SNP rankings are given including all SNPs in the MHC region and without them. Since this study uses the data from the WTCCC study, it is reasonable to expect that SNPHarvester should rank the SNPs from their results fairly highly.

Table 3. SNP hits ($p<10^{-7}$) from WTCCC study with rankings from SNPHarvester and Random Jungle.

The symbol "-" indicates that this SNP was not selected or ranked.

rsID	Chr	Position	GWAS p	RJ rank (no mhc)	SH rank (no mhc)
rs10484554	6	31382534	4.06E-214	38 (36)	14 (14)
rs240993	6	111780407	5.29E-20	184 (139)	1745 (291)
rs458017	6	111802784	2.16E-16	249 (203)	1702 (279)
rs17716942	2	162968937	1.06E-13	797 (733)	1801 (301)
rs8016947	14	34902417	1.52E-11	-	1632 (266)
rs27524	5	96127700	2.56E-11	-	1651 (271)
rs3213094	5	158683347	4.93E-11	-	1759 (298)
rs4112788	1	150817900	3.32E-10	121 (82)	1712 (282)
rs702873	2	60935046	3.59E-09	-	1884 (319)
rs6809854	3	18759427	1.12E-07	1906 (1813)	-

*4 SNP hits are not listed here because they were not replicated by SH or RJ.

2.4.2 Random Jungle

Random Jungle results consist of variable importance measures for all SNPs. There is no particular VIM value for a measure of statistical significance, but SNPs can be ranked by these VIMs.

Table 3 includes the Random Jungle VIM rankings for the SNPs previously association with psoriasis. I used the top 1000 SNPs here as I did for the SNPHarvester results, and these SNPs and their interactions were taken forward to the HyperLasso. There are several SNPs in the table which do not have rankings from Random Jungle - these SNPs were given a VIM of zero or very close to zero, indicating that the genotype data including this SNP classified individuals with the same accuracy as the permuted SNP data. In effect this means that Random Jungle did not identify these SNPs as important in distinguishing cases and controls in psoriasis.

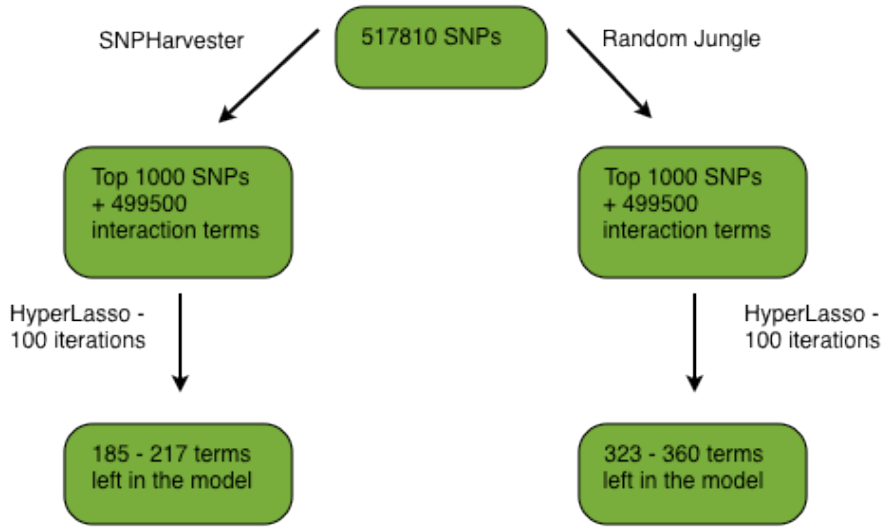

Figure 4. Workflow for SNPHarvester and Random Jungle Analyses

2.4.3 HyperLasso

The workflow and a summary of results for both the SNPHarvester and Random Jungle analyses is shown in Figure 4. Each of the 100 iterations of the HyperLasso applied to the SNPHarvester top 1000 SNPs left between 185-217 terms in the model with non-zero coefficients. After converting the SNP pairs to blocks (see section 2.3.3) many SNPs appeared in multiple iterations but only 14 terms were present in all 100 iterations. These were all interaction terms. After the Random Jungle top

1000 SNPS and their interactions were put through the HyperLasso there were between 323 - 360 terms with non-zero coefficients, with 13 present in all 100 iterations. Again these were all interaction terms. There was no overlap between these SNPHarvester and Random Jungle pairs. Table 4 shows the number of terms that were present in various numbers of iterations of the HyperLasso for SNPHarvester and Random Jungle results. There is no overlap between any of the terms from SNPHarvester and Random Jungle in this table. For example, there are 105 terms from the Random Jungle results that were selected by at least 80 iterations of the HyperLasso. Likewise there were 61 such terms from SNPHarvester that were selected by at least 80 iterations of the HyperLasso. However there were no terms that were selected by at least 80 iterations of the HyperLasso on both sets of results. In total there were 38 terms which had non-zero coefficients in at least one iteration of the HyperLasso from both the SNPHarvester and Random Jungle results, but these terms were found in far more iterations from one method than the other, and there are no obvious choices of terms that were found in many iterations from both.

Table 4. Number of terms from SNPHarvester and Random Jungle that were selected in a large number of iterations of the HyperLasso

Iteration count cut-off	Random Jungle terms	SNPHarvester terms
100	13	14
>90	66	38
>80	105	61

Our positive control - the interaction between SNPs in HLA-C and ERAP1 - was not discovered by the Random Jungle arm of the study. The VIM for the ERAP1 SNP

involved in this interaction was not ranked among the top 1000 VIMs, so the interaction between this SNP and the HLA-C SNP was not a term in the HyperLasso model. SNPHarvester did rank the ERAP1 SNP as one of the top 1000, and the HyperLasso gave the interaction term between ERAP1 and HLA-C a non-zero coefficient in 77 of 100 iterations.

2.4.4 Sparse Partitioning

To limit computational time, I required that the model selected by Sparse Partitioning should contain 4 SNPs. There are 10 possible partitionings of four predictors (allowing at most two pairwise interactions). Sparse Partitioning determined that the model with all predictors in separate groups (i.e. with no interactions) was most likely in light of the data and the prior settings (posterior probability 0.74). This model explained 46% of the variance on the observed scale. The p-values for these SNPs from the single-snp association study are given in Table 5.

Table 5. Single-SNP p-values (log-additive genetic model) for SNPs selected by Sparse Partitioning model

Chromosome	Position	SNP	Single-SNP p-value
6	31382534	*rs10484554	2.2×10^{-216}
6	31557973	rs2516509	5.265e-77
6	30048729	rs2256919	4.134e-30
5	158791476	rs10059288	1.688e-18
6	111687254	rs11153277	1.69e-10

***This SNP was given a prior probability of 1 and is not included in the 4 SNPs chosen by SP**

SNP rs2516509 is in moderate LD (r^2=0.18) with the strongly associated classical HLA allele HLAC*0602 but is not in LD with the top SNP (rs10484554). This SNP has recently been discovered as an independent signal in the MHC region which persists after conditioning on the top known MHC region hits (Knight et al., 2012). SNP rs2256919 is in almost perfect LD (r^2=0.98) with another independent MHC signal at SNP rs380924, also discovered by Knight et al. SNP rs10059288 was a previously unpublished association approximately 1kb from the gene IL12B. This SNP is not on the Perlegen platform, but rs10515802 is a perfect proxy (r^2=1.0 in CEU population of the Thousand Genomes project). This SNP had a p-value of 4.6 x 10^{-7} from the likelihood ratio test using the CASP study data. Since the analysis this SNP has been published in an analysis of psoriasis using the immunochip platform (Tsoi et al., 2012). The last association detected by Sparse Partitioning, rs11153277, is also on chromosome 6 and is in high LD (r^2=0.81) with SNP rs240993, reported by Strange et al (p=8.7 x 10^{-13}).

2.4.5 Replication of Interactions from SNPHarvester Results

SNPHarvester was the only method that selected the known interaction in psoriasis which acted as our positive control. It therefore seemed the most likely method to be finding real interaction signals. I chose to follow up on promising results from this arm of the study to try to find novel interactions in psoriasis. I identified the blocks that were present in all 100 iterations of the HyperLasso and found the SNP pairs within these blocks. Logistic regression interaction p-values from the WTCCC2 and CASP studies are given in Table 6, along with the info score from the imputation of the SNPs from the CASP study. None of the putative interaction signals suggested by SNPHarvester were significantly associated with psoriasis in the replication dataset.

Table 6. Replication of SNPHarvester interacting SNP pairs

P-values from the WTCCC2 and CASP likelihood ratio tests of interactions. Sequential rows with the same shading indicate SNP pairs from the same block pairs.

SNP Pair	SNP 1 position	SNP 2 position	WTCCC2 P-value	CASP P-value	Info Scores
rs4378174 rs3094626	1p36.3	6p22.1	0.064	0.943	0.329 0.919
rs4378174 rs3131115	1p36.3	6p21.3	0.013	0.483	0.329 0.790
rs4440829 rs6935954	1p36.3	6p22.2	0.003	0.580	0.358 0.818
rs4378174 rs434841	1p36.3	6p21.3	0.016	0.693	0.329 0.505
rs12068729 rs434841	1p36.3	6p21.3	0.004	0.731	0.336 0.505
rs10915318 rs434841	1p36.3	6p21.3	0.002	0.747	0.315 0.505
rs12128391 rs4015690	1p31.3	12q23.2	0.012	0.584	0.484 0.414
rs12402976 rs12665039	1p31.3	6p22.1	$9.29*10^{-5}$	0.847	0.512 0.825

SNP Pair	SNP 1 position	SNP 2 position	WTCCC2 P-value	CASP P-value	Info Scores
rs10084647 rs8137110	22q13.2	22q13.2	$2.70*10^{-13}$	0.760	0.303 0.212
rs7287634 rs8137110	22q13.2	22q13.2	$9.44*10^{-12}$	0.738	0.254 0.212
rs12478680 rs1416920	2q12.2	6p22.1	$1.37*10^{-4}$	0.964	0.647 0.952
rs876585 rs1416920	2q12.2	6p22.1	$5.08*10^{-4}$	0.896	0.641 0.952
rs12467322 rs4947324	2q12.2	6p21.3	0.051	0.151	0.633 0.847
rs12467322 rs3093662	2q12.2	6p21.3	.004	0.227	0.633 0.843
rs4276549 rs1131896	6p24.3	6p21.3	$3.05*10^{-4}$	0.490	0.322 0.490
rs6905949 rs1480380	6p22.1	6p21.3	$3.65*10^{-4}$	0.306	0.917 0.690
rs11752362 rs1480380	6p22.1	6p21.3	$6.8*10^{-4}$	0.170	0.957 0.690

SNP Pair	SNP 1 position	SNP 2 position	WTCCC2 P-value	CASP P-value	Info Scores
rs12663184 rs1480380	6p22.1	6p21.3	.001	0.484	0.973 0.690
rs9379859 rs9357155	6p22.2	6p21.3	0.627	0.209	0.884 0.884
rs10084647 rs6932590	22q13.2	6p22.1	0.002[3]	0.686	0.303 0.895

2.5 Discussion

The level of overlap between SNPHarvester and Random Jungle was not high enough to confidently identify promising interaction signals. If the two different methods had found the same interaction in most of the HyperLasso iterations, this would have provided more convincing evidence in support of these methods and the interactions they have found, particularly if any of the new interactions had replicated in our independent data. However this was not the case, and it was not possible to identify promising pairs of SNPs based on the degree of overlap between their results.

The SNPs in our positive control interaction (HLA-C * ERAP1) were found by SNPHarvester, and their interaction term had a non-zero coefficient in 77 of 100 iterations of the HyperLasso. This is encouraging and it seems that, for this data set,

SNPHarvester was the only method able to detect the true interaction signal in this data.

For this reason I decided to follow up the SNPs that were selected in all 100 iterations of the HyperLasso on the SNPHarvester results. However, attempts to replicate these signals in the CASP study data were unsuccessful. The SNPHarvester results could simply be false positives which will not replicate in any other data set, but the imputation quality of many of the SNPs was very poor. This would result in a loss of power to detect association signals. The CASP study was smaller than the WTCCC study, again reducing the power to detect association signals. In the absence of further, better quality genome-wide data for replication it not possible to confirm the interactions found by SNPHarvester, but it is also impossible to be certain they are not real.

Sparse partitioning identified main effects from 4 SNPs in addition to the top SNP from the GWAS, which was included as a covariate. Two of the further SNPs were shown by Knight et al to be independent associations within the MHC region. My findings indicate that Sparse Partitioning is able to identify independent signals in the same region that may be missed if study analysts simply select the top SNP in a region for follow-up. A further SNP on chromosome 6 has been previously reported as associated with psoriasis. Interestingly, Sparse Partitioning found a SNP on chromosome 5 which is not in LD with any SNP previously identified as associated with psoriasis, and the finding was replicated in the CASP data. This SNP is very near to another SNP that was identified in the WTCCC study, and it is likely that the analysts chose the top SNP in the region for follow-up. The SNP was confirmed and has now been published in another study. Although Sparse Partitioning did not identify any interactions, the SNPs it did select are sensible choices, and this method

has proven to be particularly good at identifying independent signals which may otherwise be overlooked due to proximity to other associations.

Although the methods in this study did not all replicate the findings from the original study, this is not surprising due to the different search strategies used. The interaction was originally discovered by specifically modelling interactions between all pairs of SNPs that showed single-SNP associations. This resulted in 153 tests, so an appropriate significance threshold is $0.05/153 = 3.3 \times 10^{-4}$. The interaction p-value was 2.45×10^{-5} which, while significant in this context, would not be an unexpected finding due to chance in a genome-wide interaction analysis with many more tests being carried out.

The multiple testing burden explains why the methods in this study were generally not as successful at finding this interaction, but does not explain why SNPHarvester was successful and Sparse Partitioning and Random Jungle were not. Sparse Partitioning is a processor-intensive method which requires much more computing time when more predictors are allowed to come into the model. There are many SNPs which are strongly associated with psoriasis. The study from which this data was taken listed 18 independent SNPs which were either newly-discovered or replicated in this data. All of these SNPs have a larger effect size than the interaction I was hoping to discover. Given that I had to limit Sparse Partitioning to selecting no more than 4 predictors it quite naturally selected those with stronger effects, which were all single-SNP associations. Random Jungle relies on the random selection of of SNPs available at each node in each tree. To influence the tree, one interacting SNP would have to be available at a node and be selected as the split variable over all others at that node. Then the second interacting SNP would have to be available in the daughter nodes, to allow this split to be dependent on the first SNP. All of

these less-than-certain events mean that the effect of the interaction may not have been tested frequently, and if it was it may not have been strong enough to overcome single-SNP effects or random noise in the data. SNPHarvester, unlike the other two methods, targets groups of SNPs of a specified size and tests them. Although there is a random element in terms of the starting pair of SNPs, once these are selected the algorithm then cycles through and considers all SNPs in the genome. There is a reasonable chance that SNPs which truly interact will be tested together for association, and this will increase their variable importance measure. This may give SNPHarvester an advantage in finding interactions in GWAS data.

The results of this study highlight the difficulties involved in finding statistical interactions in GWAS data. If additional genetic interaction effects exist for psoriasis, then their effect sizes must be small. Given these difficulties, I believe it is best to use multiple methods to try to find interaction effects in GWAS data, as there may be no single best method. Checking for interactions between top single-snp associations is a sensible starting point. Of the three methods compared here, SNPHarvester was the only one that had reasonable success in finding the known interaction, and therefore might be a good choice to start the search for interactions in a GWAS data set.

3 Searching for Evidence of Genetic Interactions in Protein-Protein Interactions

The search for genetic interactions in genome-wide association studies has met with limited success to date. Although significant interactions have been found in several studies using a variety of methods, they have rarely been replicated in independent data sets. The question of how frequent these hypothesized interactions actually are has still not been answered, although if they were common with large effect sizes then it is likely that more would have been found. It is possible, however, that there are many of these interactions that have not yet been found, but they have small effect sizes and are thus difficult to find with available statistical methods and sample sizes.

Finding a statistical association is simply a tool for discovering some sort of biological association that affects a particular phenotype. The belief is then that understanding the biology will help with the understanding of the phenotype. For example if the phenotype is a disease, then the genetic associations, whether they are due to a single SNP or a combination of SNPs, may highlight a biological pathway that could be a target for a treatment. In the case of multiple SNPs, a statistical interaction may be evidence that the end products created from the genetic sequences, the proteins, are interacting to affect the phenotype. While there is not very much evidence of strong statistical interactions in genome-wide association studies, there is quite a bit of publically available information on proteins that have been shown experimentally to interact with each other.

One could search for statistical interactions using biological information (e.g. pathway analysis, protein-protein interaction networks, co-expression data) to narrow the search space to genes where there is an *a priori* reason for believing there may be a biological interaction. In this study, rather than targeting areas to search for a strong association signal from specific pairs of SNPs, the focus is on looking for a general enrichment of a statistical signal from SNP-pairs within gene-pairs that code for proteins involved in known biological interactions (**P**rotein-**P**rotein **I**nteractions, or PPIs). If there is a tendency for statistical interactions to occur in these pairs, then it is sensible to hypothesize that statistical interactions do indeed occur when biological interactions are present, and the search for these interactions is a sensible way forward. If there is little evidence for a general trend for statistical interactions to occur alongside biological interactions, then perhaps the search for these interactions is not a sensible direction to take. Instead resources could be directed to increasing sample sizes in single-locus studies, or using sequencing technology to find rare variants.

I have created a software tool which starts with a list of PPIs, a map of SNPs to genes coding for these proteins and GWAS data. The output from this program enables the computation of a pair of test statistics which summarize the p-values from SNPs in genes coding for interacting proteins, along with a null distribution created by permuting the input data.

3.1 Statistical Methods

The analysis is carried out in several stages. First, all pair-wise interaction p-values are calculated for all input SNPs. This includes the pairs of SNPs from the paired genes, but also all other combinations of SNP pairs. Calculating the p-values from

SNP pairs which are not interacting will allow the creation of a null distribution for our test statistic. Next, two test statistics are calculated by summarizing selected p-values from SNPs within the genes in the paired gene list. Finally, a permutation scheme runs a user-defined number of permutations which allows the creation of null distributions for both test statistics. Each of these stages is described in detail in the following sections, and the outline algorithm is given in Figure 6.

3.1.1 Interaction p-values

After creating our list of SNPs for analysis we calculated a p-value for interactions using PLINK's fast-epistasis command. This carries out a test for differences in odds ratios between cases and controls. The interaction coefficient β in a logistic regression model is the same as the ratio of the odds ratios of the cases and controls.

Equation 8. Interaction coefficient in logistic regression

$$\beta = \log(R/S) = \log(R) - \log(S)$$

If the two odds ratios are the same, then R/S=1and β=log(1)=0, indicating that there is no interaction between the two loci. PLINK's implementation of this test is given in Equation 9.

Equation 9. Plink's fast-epistasis test

$$Z = (\log(R) - \log(S)) \; / \; \mathrm{sqrt}(SE(R) + SE(S))$$

where R is the odds ratio of the two alleles in cases and S is the odds ratio of the two alleles in controls. This follows a standard normal distribution.

PLINK's fast-epistasis command is not quite equivalent to the test of the logistic regression coefficient since the phasing of the alleles between the two loci is not known, as it would be in the equivalent logistic regression model. PLINK's odds

ratios are calculated simply with allele counts. However the authors claim that the correlation between PLINK's test and logistic regression is very high (r=0.995).

3.1.2 Test Statistics

The purpose of the test statistic in this study is to summarise the interaction p-values for SNP pairs from gene pairs involved in PPIs and compare them to p-values from SNP pairs that aren't involved in PPIs. I have implemented the calculation of two test statistics that might be useful for this comparison, one based on a summation and the other a proportion.

The first test statistic involves summing a transformation of the p-values across all SNPs within genes coding for known interacting proteins (see Equation 10).

Equation 10. PERSI test statistic: sum of interaction p-values

$$T_1 = \sum_{i=1}^{n_{int}} (-2 * \ln(p_i))$$

where:

n_{int} is the number of SNP-pair p-values from interacting gene pairs

p_i is the p-value for the interaction from the *i*th interacting SNP pair.

An alternative to simply summing p-values, this transformation makes the meaning of the sum more intuitive, so that the higher the value of the test statistic, the more significant the results. The total sum is then adjusted by dividing by the number of values summed, as this can vary depending on the number of SNPs in the genes that are included. It can be shown (Fisher, 1925) that T_1 follows a chi-square distribution with $2*n_{int}$ degrees of freedom under the null if all p-values were obtained from independent tests. Thus $E[T_1/n_{int}]=2$ regardless of the number of SNPs. In our case,

the p-values are unlikely to be independent but it is still reasonable to suppose that $E[T_1/n_{int}]$ is a constant.

Although we are searching for an enrichment of low p-values, it is not expected that the number of interacting pairs will be a large proportion of the total that are considered. Because of this there is a large amount of 'noise' in the data, because the likelihood of any one SNP pair being important is low. In order to reduce this noise, we applied this same summation to only SNP pairs whose p-values met certain user-defined thresholds.

Equation 11. PERSI test statistic: sum of transformed p-values below user-defined threshold

$$T_1 = \sum_{i=1}^{n_{int(p1)}} (-2 * \ln(p_i))$$

where:

$n_{int(p1)}$ is the number of SNP-pair p-values from interacting gene pairs that are below threshold p1

If there is an enrichment of low p-values within our target SNP pairs, then the test statistic based on the sum of the transformed p-values will be higher than that obtained by summing p-values from other SNP pairs.

The second test statistic was the proportion of 'significant' p-values between PPI SNPs, again at various user-defined thresholds. If statistical interactions can be found from SNP-pairs within genes coding for protein-protein interactions then a higher proportion of low p-values would be expected than from SNP-pairs from other genes.

Equation 12. PERSI test statistic: proportion of low p-values between SNPs in PPIs

$$T_2 = \frac{n_{int(p1)}}{n_{int}}$$

As with T_1, this value should be higher if there are more interacting SNP pairs from the gene pairs in the PPI list than from other pairs of SNPs. This proportion is then compared to the null distribution generated from the permutation stage.

3.1.3 Permutation scheme

To assess the importance of the test statistics we devised a permutation scheme to create a null distribution and compared our test statistic to this to calculate an empirical p-value. In order to preserve the genetic structure of the data (for example, variable gene sizes and linkage disequilibrium) we permuted only the gene labels within the gene-pair list and recalculated our two statistics in a manner similar to a Mantel test (Mantel, 1967) (see Figure 5).

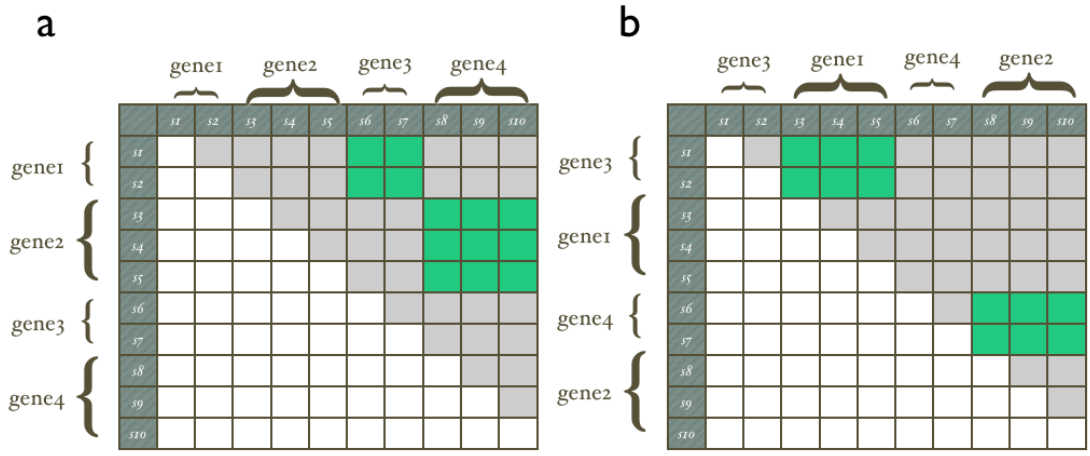

Figure 5. Permutations scheme for gene labels

Green highlighting indicates the interacting gene pairs (1 & 3, 2 & 4) used to calculate the test statistic. The real gene labels are given in a), while in b) the gene labels are permuted but the SNP labels stay the same, preserving gene structure. The green areas in a) will contribute to the real test statistic, while the green areas in b) will contribute to a value from the null distribution.

The list of SNPs belonging to a single gene remained the same, but the identifier of the gene changed. The list of interacting gene pairs also remained the same. As a

reminder, for both statistics a higher than expected value for the real test statistic will indicate an enrichment of low p-values in the SNPs involved in PPIs.

The algorithm for the whole process is given in Figure 6.

Inputs:

D: genotype data for N_a cases and N_u controls

PVALS: a set of *M* p-value thresholds

PPI: a paired gene list

S2G: a SNP-to-gene map

NPERM: number of permutations

RESULTS: name for results file

Output:

1. Simplified interaction p-values file for further analyses
2. Results file with counts and summed -2*ln(p) for each threshold for real data (first row) and permuted data (next NPERMS rows)

Create set list SNPSET of all unique SNPs in S2G

Create gene list GENELIST from all unique genes in S2G

Call PLINK to calculate interaction p-values from D, between all SNPs in SNPSET and output results to file SNPINTS

#Simplify each result to an integer-format gene-pair ID and associated p-value

Read in SNPINTS, S2G, GENELIST

For i in 1:length(SNPINTS)

Replace pair of SNP identifiers with gene pair identifier:

Look up corresponding genes for both SNPs in S2G

Get position indices for these genes from GENELIST

Combine GENELIST indices using the formula (m+1)*NGENES + (n+1), where m is the smaller index and n is the larger

Resulting integer is a unique PAIRID for this gene pair

Output PAIRID and current p-value to the file GENEINTS_SIMPLIFIED

#Calculate test statistics for each p-value threshold, carry out permutations

Read in GENEINTS_SIMPLIFIED, PPI, PVALS

For i in 1:(NPERM + 1)

Count total number of p-values for PAIRIDs corresponding to gene pairs in PPI and store in PCOUNT.ALL

Sum -2*ln(p) for all p-values for PAIRIDs corresponding to gene pairs in PPI and store in PSUM.ALL

Output to RESULTS (line i) as space-delimited: PAIRID PCOUNT.ALL PSUM.ALL

For j in 1: length(PVALS)

Count p-values below PVAL(j), store in PCOUNT.j

Sum -2*ln(p) below PVAL(j) and store in PSUM.j

Output to RESULTS (line i) as space-delimited: PCOUNT.j, PSUM.j

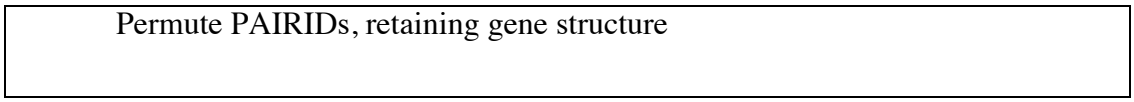

Figure 6. PERSI algorithm.

The process is managed by a bash shell script. It calls PLINK to calculate interaction p-values, and two C++ programs to 1) simplify PLINK results and 2) calculate statistics and permute data. The simplification step can be skipped if the simplified results file is already available from a previous analysis.

3.2 Application to WTCCC Crohn's data

The Wellcome Trust Case Control Consortium (WTCCC) carried out a genome-wide association study in Crohn's disease (Wellcome Trust Case Control Consortium, 2007), and this data is freely available to researchers. Genotyping was done on the Affymetrix GeneChip Mapping 500k assay.

3.2.1 Quality Control

Data from the WTCCC study is available separately for each of 7 phenotypes and 2 sets of unscreened population controls. For this study I used the Crohn's disease cases (CD), National Blood Service (NBS) and 1958 birth cohort (58BC) controls, all of which were genotyped at 500568 SNPs. There were 2009 CD, 1500 NBS and 1504 58BC individuals genotyped. Each of the three data sets came with a list of SNPs and individuals that failed quality control (QC) carried out by the WTCCC. As an extra check I merged the three files together first and then carried out QC procedures, to ensure that the merged data set was as clean and high-quality as possible. I used the same thresholds used by the WTCCC in their QC (see Table 7), and I removed individuals before carrying out SNP level QC. Due to the two control sets being genotyped separately it was sensible to check for significant differences in allele frequencies between them. I carried out a pseudo-association study using NBS/58BC status as case/control status, and excluded SNPs on the basis of highly

significant p-values. Summaries of the individuals (Table 8) and SNPs (Table 9) removed by my QC and the WTCCC QC are given below.

After carrying out my QC as described I then ensured that all SNPs and individuals in the WTCCC exclusion lists were removed. As a result, all SNPs or individuals who were problematic in either the separate data files or the combined data set were removed from the analysis.

Table 7. Quality control thresholds for merged WTCCC Crohn's data and controls

QC metric		Threshold
Individual	Missingness	3%
	Sex check	Phenotype/genetic mismatch or 0.2<F<0.8
	Heterozygosity	<0.225 or >0.3
SNP	Missingness (Miss) and Minor Allele Frequency (MAF)	Miss>0.05 OR Miss>0.01 AND MAF < 0.05
	HWE	$5.7*10^{-7}$
	Trend (NBS vs 58BC)	$5.7*10^{-7}$
	X-chromosome	All

Table 8. Individuals removed by author's QC, WTCCC QC and combined.

		Number removed		
Reason	**Sample Origin**	**WTCCC**	**Author's QC**	**Combined (unique)**
Missingness > 3%	NBS	8	8	8
	58BC	9	9	9
	CD	44	48	48
	All miss	**61**	**65**	**65**

Sexcheck	NBS	0	2	3
	58BC	0	3	2
	CD	0	1	1
	All sex	**0**	**6**	***6**
Heterozygosity	NBS	0	0	0
	58BC	0	0	0
	CD	4	4	4
	All het	**4**	**4**	****4**
***Other reasons	NBS	34	0	34
	58BC	15	0	15
	CD	209	0	209
	All others	**258**	**0**	**258**

Grand totals	**NBS**	**42**	**10**	**45**
	58BC	**24**	**12**	**26**
	CD	**257**	**53**	**262**
	All cohorts	**323**	**75**	**333**

***Author's sex check found 15 individuals – 9 of these were included in the WTCCC removals for 'other reasons', and are not included here.**

****Author's heterozygosity check found 9 individuals – 5 were included in the WTCCC removals for 'other reasons' and are not included here.**

*****Other reasons for exclusion by WTCCC- Affy CEL file bad, external discordance (DIL genotyping, NBS blood type, phenotype measure in CD), non-European ancestry, sent twice by Sanger, duplicated at unknown stage, 1st/2nd degree relatives**

Table 9. SNPs removed by author's QC (after merging 3 files) and WTCCC QC (performed separately on each file).

Reason	Author's QC	WTCCC
Missing/MAF	24910	26190
HWE	4381	4303
Trend	103	93
X-chr	10152	370
Total	39456	30956

After quality control measures were applied, there were 458210 SNPs genotyped in 1747 Crohn's disease cases and 2933 controls.

3.2.2 Protein-protein interactions and SNP-gene mapping

Protein-protein interactions were identified by a bioinformatician colleague Benjamin Lehne, using information from 5 manually curated databases according to his previously described work (Lehne and Schlitt, 2009). In order to be as certain as possible that we were including only true PPIs, only those that had been confirmed in

at least two independent publications were used. All interactions that arose from well-known protein complexes as defined by MIPS CORUM (downloaded 3.06.2008) were removed. All genes that had an interaction in the above set were mapped to the WTCCC Crohn's disease study SNPs. For mapping SNPs to genes, a SNP had to lie within a gene or 5kb to either side of the gene. SNPs that could not be assigned unambiguously (for example, they were within mapping distance to more than one gene) were removed.

The above exercise resulted in

- 1015 protein-protein interactions
- 1165 genes coding for these proteins
- 15010 SNPs from Affy 500 mapped to these proteins
- 112,642,545 possible SNP pairs
- 86,231,572 pairs after excluding SNPs for QC.

Our first test statistic, the sum of the transformed p-values, was calculated for all SNPs in the PPIs as well as including only p-values below the thresholds 0.10, 0.05, 0.01 and 0.005. Our second test statistic, the proportion of p-values below a threshold, was calculated at p-value thresholds 0.05, 0.01 and 0.005.

3.2.3 Results

The first stage of PERSI produces p-values for all pairwise comparisons of SNPs given to PLINK in the SNP set list. This represents a set of SNPs in which there is some reason to believe there may be biological interactions. It is therefore sensible to check the individual p-values for these SNP pairs. The most significant p-values from the PLINK fast-epistasis analysis are given in Table 10. Given that we are testing 86 231 572 SNP pairs, a reasonable Bonferroni-corrected level of significance

is 0.05/86231572 =5.8*10^{-10}. There are no SNP pairs with a p-value lower than this, so we have not found any interacting SNP pairs that are significantly associated with Crohn's disease.

Table 11 contains the results from the analysis using all of the p-value thresholds we applied. There is no evidence here to support our hypothesis that there is an enrichment of low p-values within genes coding for proteins involved in protein-protein interactions.

Table 10. Top interaction p-values for PLINK's fast-epistasis analysis on SNPs in genes coding for proteins involved in protein-protein interactions

SNP1	SNP2	p-value
rs10276324	rs17570793	1.84E-08
rs11208830	rs12108497	3.17E-08
rs3845731	rs17536147	5.26E-08
rs17507748	rs2919387	7.26E-08
rs17507685	rs2919387	9.60E-08
rs2918417	rs12934281	9.74E-08
rs1934909	rs500629	1.08E-07
rs7894595	rs8114927	1.31E-07
rs2918417	rs7196708	1.54E-07
rs2918417	rs11647877	1.71E-07

Table 11. Empirical p-values for two test statistics at multiple p-value thresholds.

	Empirical p-value	
p-value threshold	**∑-2ln(p)**	**proportion**
all results	0.77	-
p<.05	0.08	0.28
p<.01	0.10	0.12
p<.005	0.23	0.13

3.3 Discussion

None of our tests have provided convincing evidence of an enrichment of low p-values in SNP pairs corresponding PPIs in Crohn's disease. There are hints of enrichment at some thresholds, but it is difficult to hypothesize as to why these particular thresholds should be important, and we certainly can't rule out that these modestly low p-values are due to chance.

One limitation of this study is the application in just one phenotype. There is no particular reason to believe that PPIs are important in Crohn's. It was selected for convenience, as one of many phenotypes for which there is now publically available GWAS data.

The genotyping in the WTCCC Crohn's study was done on the Affy500k chip, but more modern chips offer better coverage of the genome. Denser, better-designed chips may be more successful in tagging or typing causal variants, which will increase the probability of finding both single-SNP effects and interactions, since the power to detect statistical effects will decrease with imperfect tagging.

PERSI is designed to look for an enrichment of low p-values within gene pairs, but this will not necessarily be present if a phenotype is strongly influenced by a single SNP pair, or a single pair of genes. In this case the statistical noise from all of the pairs that are not associated with the phenotype, but are in our list of reported interacting pairs, could mask the signal from the true interacting SNPs. In this case other methods for detecting SNP-SNP interactions would be more useful. Examining the interaction p-values from all of the SNP pairs is an obvious starting point, but in this particular study there were no particularly low p-values.

Our processes of selecting PPIs and assigning SNPs to genes could be further areas where a different approach could provide different results. Any list of gene pairs could be used as an input to PERSI, and we have only used one approach of using well-published PPIs. It is also possible to limit the analysis to gene pairs within pathways that have been implicated in a particular phenotype, potentially reducing statistical noise from the amount of testing done with SNPs that are unlikely to be involved. The assignment of SNPs to genes could also be done using different methods, for example larger or smaller flanking regions around a gene.

Investigation of further phenotypes using different GWAS data, gene pair lists and SNP-to-gene assignment methods may yet provide evidence of an enrichment of interaction signals.

4 Genome-Wide Association Study in Renal Transplantation

In this chapter I will describe my contribution to a GWAS in renal transplantation that was carried out at King's College London.

4.1 Background

The aim of this project was to find genetic variants associated with the survival time of a transplanted kidney, as well as a number of secondary phenotypes. The design of this study is somewhat different to a standard GWAS as the unit of analysis for the primary phenotype is a transplant, not an individual. Genotypes from two individuals contribute to each transplant event, and it is plausible that genetic variants in donors and recipients could both affect the prognosis for the transplant.

My contribution to this study involved work on a section of the quality control (QC) procedures as well as the analysis of the binary phenotypes. I carried out a principal components analysis to identify non-Caucasian samples as part of the QC. I then carried out the analysis of three binary phenotypes; progression to end-stage renal failure, acute regression and intra-cranial haemorrhage. I worked on the analysis of the main study data collected at King's College London (KCL) as well as a replication cohort to follow up the most likely SNPs from each phenotype.

4.2 Genetic Data and Phenotypes

Genotyping was carried out in-house at KCL on Illumina Quad 610 chips. Donors and recipients for 625 transplants (1250 samples) were genotyped. One thousand one hundred and fifty-two samples (576 complete transplants) passed QC.

The primary phenotype for this analysis is time to graft failure in days, however in this study I analysed three binary phenotypes in a standard case/control GWAS design. Details of the survival analysis and results that I carried out on an expanded data set are presented in Chapter 5.

4.2.1 End-Stage Renal Failure

All of the recipients of a kidney transplant have progressed to end-stage renal failure (ESRF). Although the underlying disease may be different, all of these patients' kidneys have failed, while other patients with the same diseases will not progress to ESRF. The hypothesis behind this analysis is that there may be an underlying genetic cause for progression to ESRF, which could be due to a donor or recipient variant or an interaction between the two. I therefore used all 576 recipients of kidney transplants as cases in this case/control analysis. Various sets of controls were used, described below.

4.2.2 Acute Rejection

An acute rejection is often defined by sharply raised creatinine levels in blood, but for a definitive diagnosis a biopsy is examined for histological changes. For this study only biopsy-confirmed acute rejection episodes were used as cases, and all kidney transplant recipients who had 0 recorded AR episodes were used as controls. Using this information, there were 173 cases and 356 controls for the case/control analysis of acute rejection.

4.2.3 Intracranial Haemorrhage

A large proportion of transplant donors in the UK are otherwise healthy victims of a sudden stroke. Donor records indicate the cause of death for deceased donors, and a

large number of organ donors have a cause of death of intracranial haemorrhage (ICH). This information allowed us to carry out a GWAS in ICH using the donor samples that have already genotyped.

4.2.4 Control Samples

For the case/control analyses of ICH and ESRF, I initially used freely available, unscreened controls from the Wellcome Trust Case Control Consortium Phase 1. These consisted of 1500 samples from the 1958 British birth cohort (58BC) which were genotyped on the Illumina HumanMap 550k chip. This is a less dense chip than the Illumina Human610-Quad which was used for the transplant patient samples, so there were some SNPs that did not have control genotypes and could not be analysed.

In order to increase sample size of controls, and therefore power, and improve SNP coverage to match the transplant study genotypes, I carried out the same analysis using the WTCCC Phase 2 controls. These controls consisted of 3000 58BC and 3000 NBS unscreened individuals, each of which consisted of the original 1500 controls genotyped for Phase 1 along with 1500 new samples. These samples were genotyped on the Illumina Human670-Quad chip. This chip is a slight upgrade to the Illumina Human610-Quad used for the KCL transplant patients, so the overlap was very high and fewer SNPs were lost from the analysis.

Table 12. Data available for ESRF and ICH association studies

Control Source	ESRF	ICH
WTCCC1	Cases: 587 Controls: 1430 SNPs: 448934	Cases: 226 Controls: 1430 SNPs: 448934
WTCCC2	Cases: 587 Controls: 5013 SNPs: 546112	Cases: 226 Controls: 5013 SNPs: 546112

4.2.5 Replication

Replication data was obtained from a cohort collected and genotyped by a team at the University of Newcastle. The Newcastle (NC) cohort consists of 400 donors and 439 recipients. The 439 recipients were used as ESRF cases for replication, and the 400 donors were used as controls. One hundred and fifty-three of the donors died of an intracranial haemorrhage, and these were used as cases for ICH replication with the remaining donors as controls.

4.3 Methods

4.3.1 Quality Control - Principal Components Analysis

My contribution to the quality control procedures in this study was to carry out a principal components analysis (PCA) to identify non-Caucasian samples. I received a data set that had been partially checked for quality control measures using standard

QC techniques. Thresholds applied for SNPs were missingness greater than 2%, HWE p-values less than 10^{-4} and minor allele frequencies less than 1%, resulting in the removal of 67748 SNPs. Thirty-eight individuals were removed for sample missingness (5%), phenotype/genotype sex discordance and duplicated samples.

After QC had been carried out to this stage the data set was pruned using linkage disequilibrium (LD) to a smaller set of SNPs. For the purposes of genetically distinguishing between ethnic groups a reduced set of SNPs is adequate, and this reduction is generally done by thinning out the data to a set of SNPs with low LD. For this work I was given the file with QC applied as above, and LD-pruned using PLINK. This pruning process in PLINK compares SNPs within a window of a user-specified size, which in this case was 1500BP. After checking within the first 1500BP of a chromosome the window 'slides' along by another user-specified amount- in this case 150 BP - and checks LD again. This sliding window continues along until the end of the chromosome. The r^2 threshold used for this data was 0.20, so one SNP from each pair of SNPs that are within 1500BP of each other, and with an r^2 of greater than 0.20, was removed. The resulting LD-pruned file had 80760 SNPs in it.

In order to identify samples from non-Caucasian populations I merged the renal transplant samples with data from the HapMap project (International HapMap Consortium, 2003). I used all available unrelated samples from the CEU (Utah residents of European ancestry), YRI (Yorubans from Ibadan, Nigeria), JPT (Japanese from Tokyo, Japan) , CHB (Han Chinese from Beijing) and GIH (Gujarati Indians in Houston, Texas) collections as genotyped in HapMap Phase 3.

Table 13. HapMap samples used in PCA for population stratification

Population	Sample Size
CEU	112
YRI	113
JPT	86
CHB	84
GIH	88

The data sets were merged only on the SNPs that were common between all HapMap samples and the renal transplant samples. This left a total of 58035 SNPs for the PCA.

The PCA was run in R and eigenvalues and eigenvectors were extracted using the *eigen()* function. Given that there are 4 ethnic populations in our data, the majority of the genetic differences due to population differences should be accounted for in the first 3 PC axes. The plots in Figure 7 illustrate the ethnic group to which the renal samples (light blue) belong by plotting PC1 vs PC2 values (top) and PC1 vs PC3 values (bottom). Each dot on this plot represents a single sample. A small number of the renal samples (light blue dots) clearly overlap or move towards a population other than CEU. The CEU samples are difficult to see but they are black circles clustered under the majority of the renal samples near the bottom-centre of the first plot and bottom-left of the second.

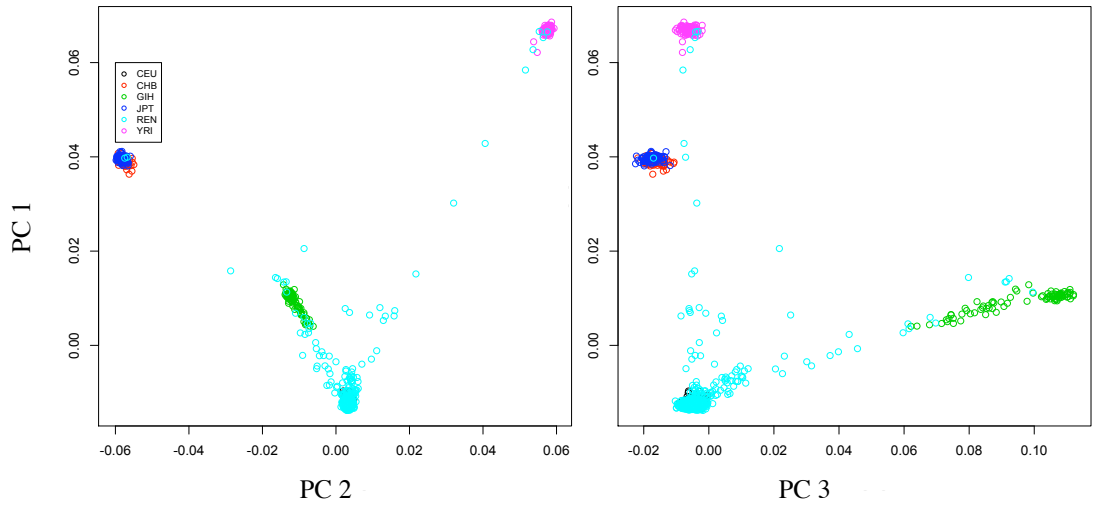

Figure 7. First 3 PC axes of PCA for population stratification

The aim of this PCA was to identify and remove samples from people who were not Caucasian, and this was done by setting thresholds for PC1, PC2 and PC3 that tightly surrounded the CEU samples, and the renal samples which overlap them. Any renal samples which did not meet these thresholds were then removed from further analysis. I set thresholds of PCA1 scores < 0.01 and PCA3 scores <0.02 for inclusion in the GWAS. Thirty-two samples did not meet these criteria and were removed for having non-Caucasian ancestry.

4.3.2 End Stage Renal Failure, Acute Rejection and Intracranial Haemorrhage

These three studies were all carried out as standard case/control analyses using PLINK. This performs a chi-square test of association between allele counts at each SNP and case/control status. The results consist of p-values for each SNP indicating the significance of the finding, along with odds ratios for the 2x2 contingency table.

4.4 Results

4.4.1 End Stage Renal Failure

On analysis with the WTCCC1 controls, two of the ESRF p-values were far smaller than any others (rs10053502, p-value=1.86E-51 and rs361147, p-value=5.11E-38, see Figure 8 top plot). No other SNPs near these two showed evidence of association, which is normally seen because of LD. On contacting Illumina we found that some SNPs showed large differences in allele frequencies when genotyped on the Human610-Quad chip (renal samples) and the HumanMap 550k chip (WTCCC1 controls).

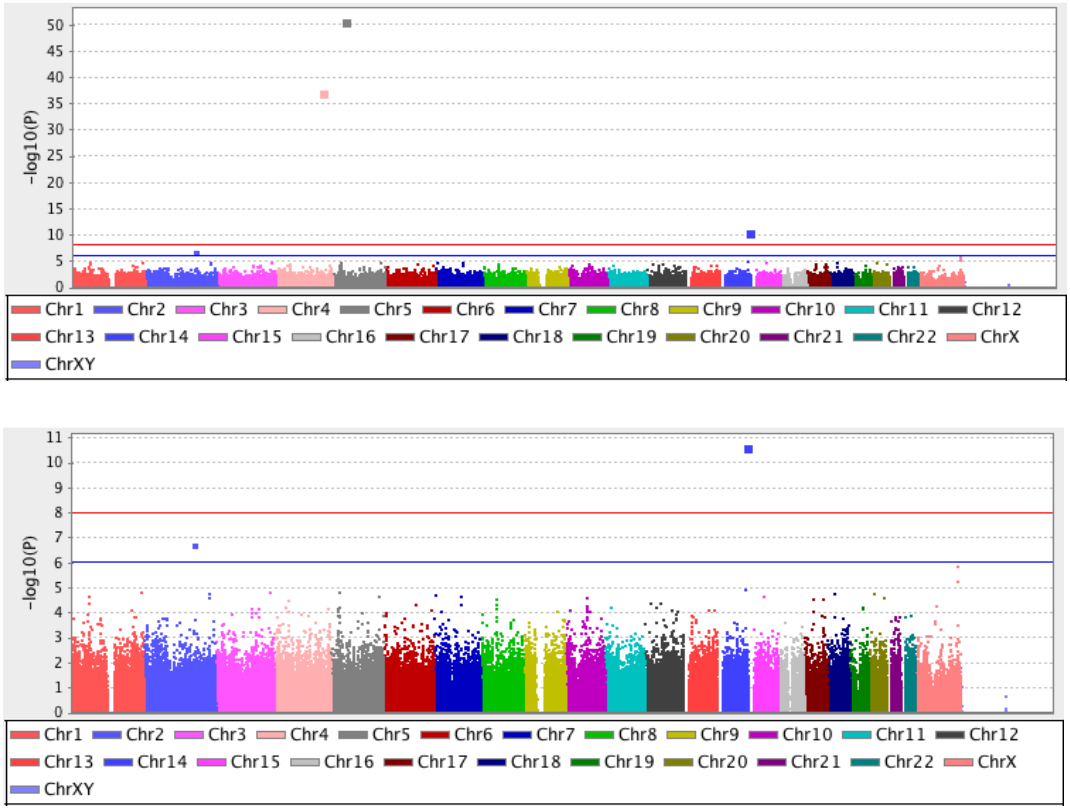

Figure 8 Plots of p-values from ESRF analysis using WTCCC1 controls

Results shown before (top) and after (bottom) removal of 2 SNPs for confirmed platform discrepancies. Note different scale on the y-axis.

We obtained a list of these SNPs and our top two were in this list. We therefore removed these two SNPs and re-examined the results (see Figure 8 bottom plot).

The results of the association study using the larger set of WTCCC2 controls are given in Figure 10. The top figure shows a cloud of SNPs above the majority of the plotted SNPs which have very small p-values. However, as with the top two SNPs in the analysis with the WTCCC1 controls, there are no neighbouring SNPs showing association signals and thus no 'towers' common with real association signals. Given the very low p-values for the SNPs in this cloud, it would be very unusual if no other SNPs in the region of any of these 30 SNPs showed any association signal at all. These SNPs were not on the list of SNPs given to us by Illumina, however we believe these to be false signals due to differences in the platforms (610Quad and 670Quad), and we relayed this information to Illumina. A corrected plot with this cloud of SNPs removed is also given in Figure 10.

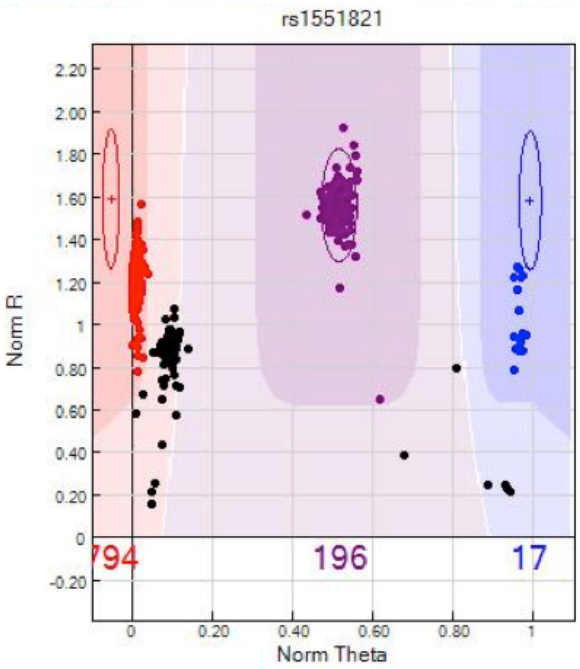

Figure 9. Cluster plot of rs1551821 showing poor clustering an genotype calling

The remaining top SNP association on chromosome 18 was an isolated SNP showing no 'tower' of nearby associations. Singleton associations such as this raise concerns over genotyping. Examination of the cluster plot (see Figure 9) showed that this SNP had not been called properly, and this SNP was not replicated.

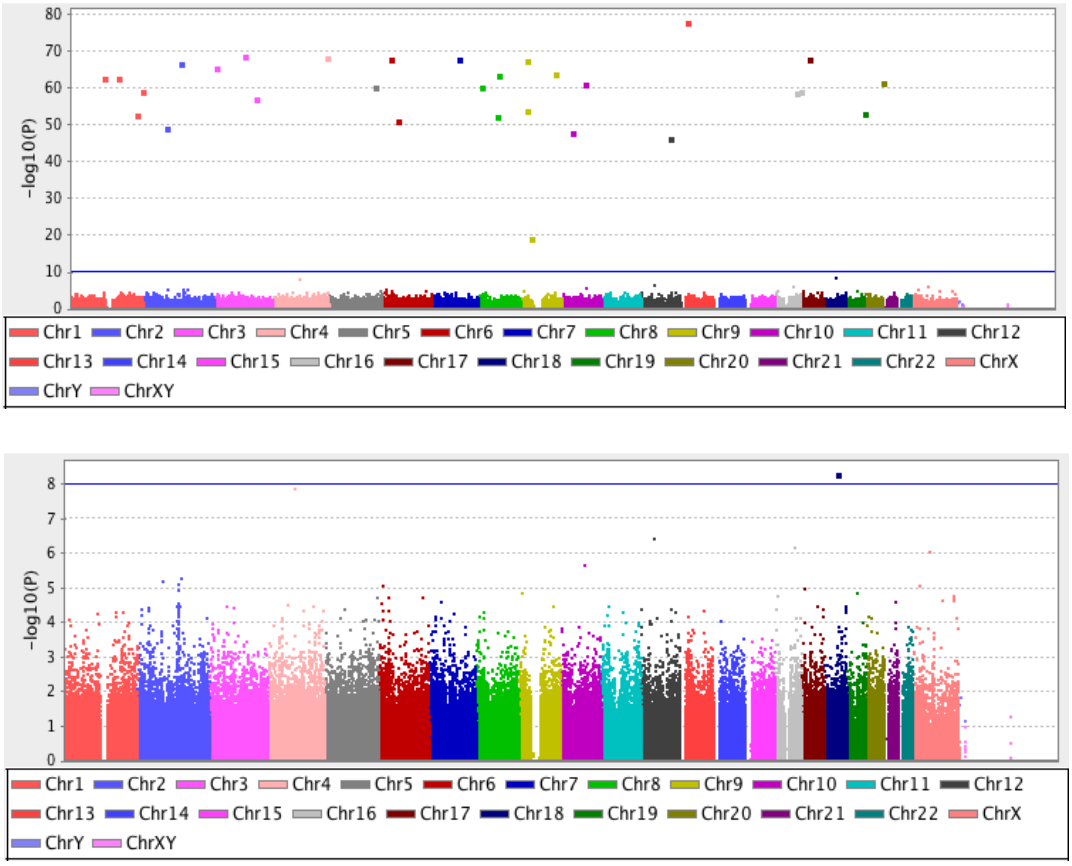

Figure 10. Plots of p-values from ESRF analysis using WTCCC2 controls

Results shown before (top) and after (bottom) removal of 2 SNPs for confirmed platform discrepancies. Note different scale on the y-axis.

After removing the SNPs that were believed to be false association signals due to platform differences and the SNP that was poorly genotyped, the remaining top SNPs from the analysis with WTCCC1 and WTCCC2 controls were genotyped in the Newcastle replication samples. The results of the replication analyses using Newcastle recipients as cases and donors as controls are given in Table 14.

Table 14. Association (KCL vs WTCCC controls) and replication (Newcastle recipients vs donors) p-values for ESRF

SNP	Chr	Position	KCL P-value	Newcastle P-value
rs874838	2	85275202	6.52E-06	0.3729
rs2289959	2	136140374	8.08E-06	0.7278
rs309152	2	136373722	8.10E-06	0.3239
rs16829355	2	149306740	5.59E-06	0.8894
rs950266	3	57324905	3.82E-06	0.8366
rs4484262	4	89158024	1.32E-08	0.9397
rs10014233	4	186017111	5.00E-07	0.6108
rs2786714	9	103624285	5.60E-06	0.3383
rs704018	10	80291027	2.33E-06	0.2051
rs10500537	16	65160772	6.73E-07	0.5028
rs8074821	17	71804694	4.19E-05	0.6252
rs7059717	X	54436892	9.18E-07	0.095

4.4.2 Acute Rejection

The Manhattan plot of recipients with biopsy-confirmed definite AR vs recipients with unconfirmed AR or no record of AR is given in Figure 11.

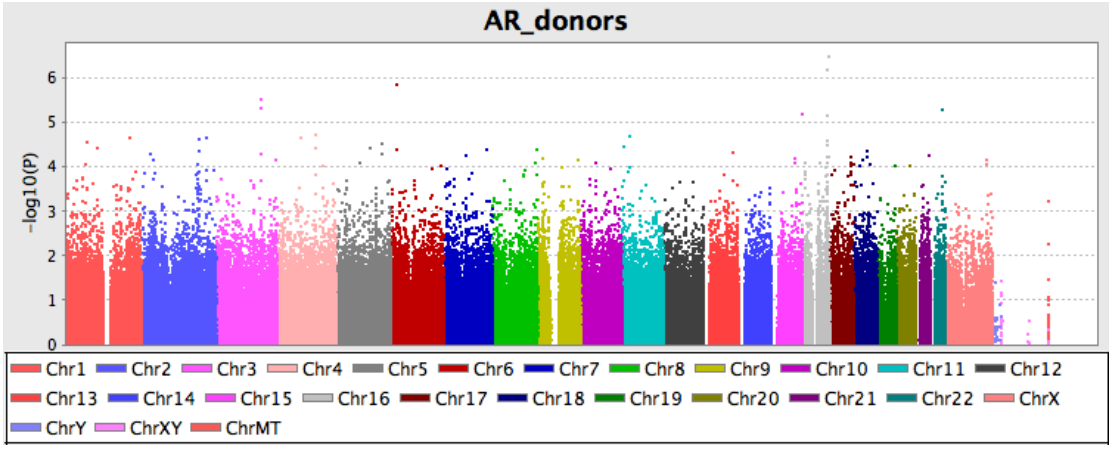

Figure 11 Association p-values for KCL recipients with at least one acute rejection episode in the first twelve months post-transplantation

Data uses KCL recipients who did not have an AR as controls.

Table 15 P-values of top associations for Acute Rejection

SNP	Chr	Position	P-value
rs7431637	3	144532459	4.83E-06
rs4839628	3	144569015	3.08E-06
rs17051344	4	122504713	1.88E-05
rs3759	6	15405889	1.42E-06
rs7951530	11	24203480	2.11E-05
rs12324075	15	98690270	6.63E-06
rs1401197	16	78895134	6.75E-07

SNP	Chr	Position	P-value
rs3098599	16	79371197	7.17E-06
rs11865270	16	87183002	3.41E-07
rs114607	22	38706329	5.29E-06

Acute rejection phenotype information was not available at the time of analysis for the Newcastle recipients. This data is available from NHSBT, however we did not ask for it at this stage as we asked for all phenotype and covariate data for all UK transplants at a later date, for the WTCCC3 study described in the next chapter. As a result replication of these associations was not attempted at the time of analysis.

4.4.3 Intracranial Haemorrhage

For the analysis of donors who died of ICH, we used the same WTCCC1 and WTCCC2 controls as we used in the ESRF study. The platform differences are again apparent, giving spurious association signals for 2 SNPs using the WTCCC1 controls and 29 SNPs using WTCCC2 controls. Plots of the p-values from these analyses, before and after removal of the spurious associations, are given in Figure 12 and Figure 13.

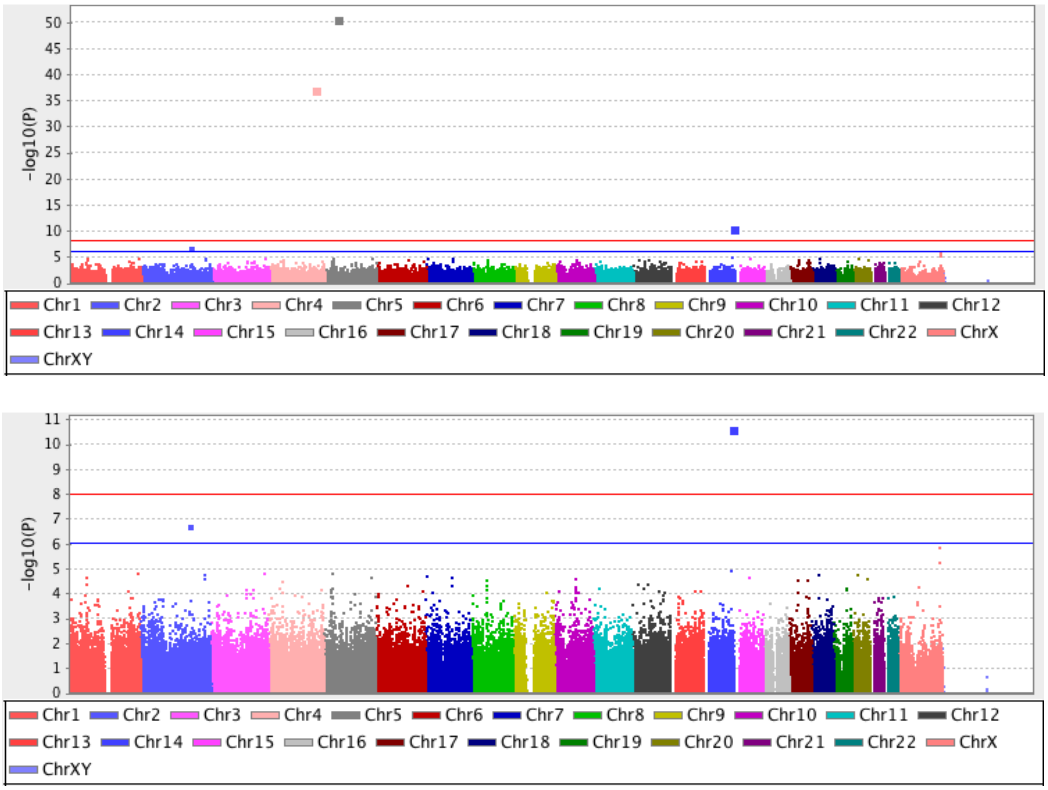

Figure 12 Association p-values for ICH using WTCCC1 controls

Data represents before (top) and after (bottom) removal of SNPs showing spurious association signals due to platform differences.

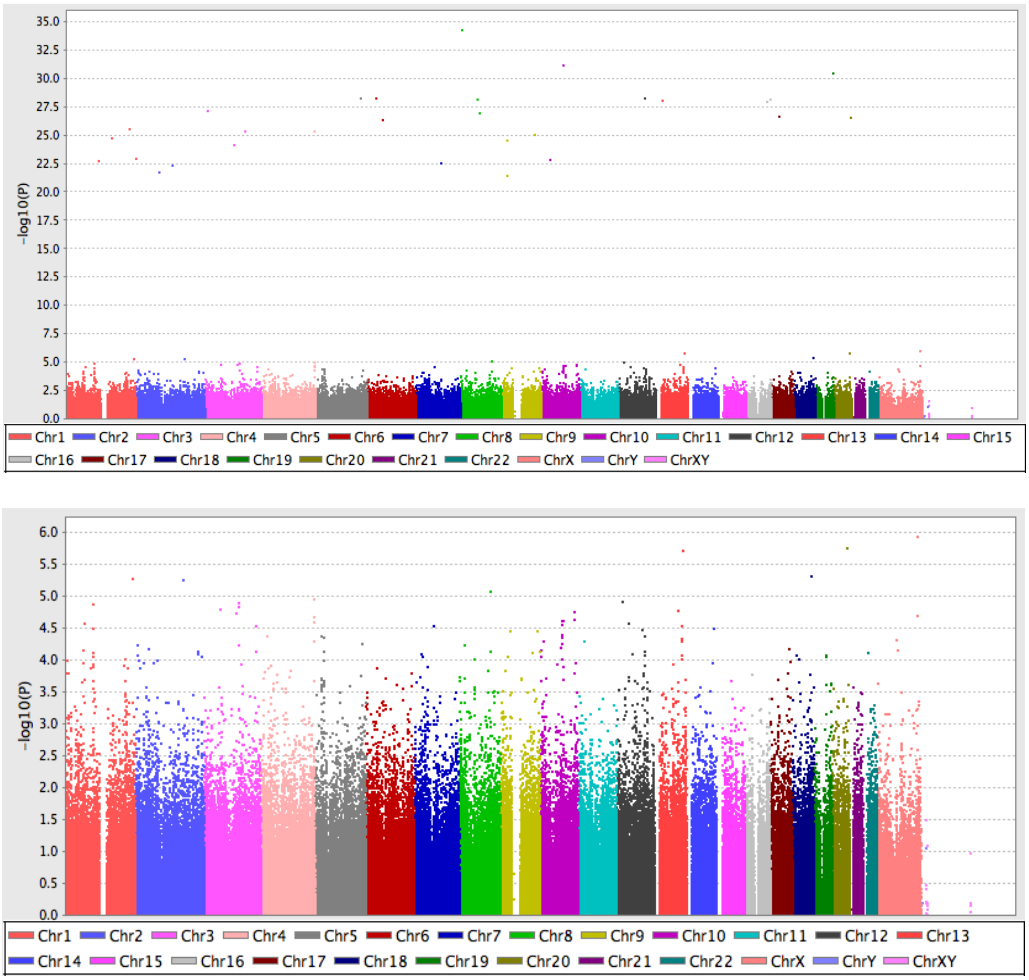

Figure 13 Association p-values for ICH using WTCCC2 donors

Data represents associations before (top) and after (bottom) removal of spurious association signals.

The cloud of spurious associations from the WTCCC2 controls is 30 SNPs in the ESRF study, and 29 of those form the cloud in the ICH study. The 30th SNP in the ESRF cloud is rs855523 on chromosome 9. Although the p-value for this SNP in ICH is not in our cloud, it still a very small p-value at 3.543×10^{-5}. It is unlikely that this SNP would have such a low p-value for both of these phenotypes, so we assumed that this was also due to genotyping problems and did not attempt to replicate it.

Replication of the top hits from the ICH study was unsuccessful. The power of the replication study is very low; even lower than for ESRF, with only 153 cases of ICH. We can't conclude whether or not the top hits in the discovery analysis are false positives or true associations that didn't replicate due to lack of power in replication.

Table 16. ICH SNPs with discovery and replication p-values

SNP	Chr	Position	Discovery p-values	Replication p-value
rs9428851	1	238790962	5.38E-06	0.3041
rs6714431	2	168885012	5.79E-06	0.0692
rs9854823	3	121633974	1.29E-05	0.5158
rs1288569	4	186467270	1.15E-05	0.0783
rs1283726	8	108545032	8.46E-06	0.8106
rs2227564	10	75343107	9.89E-05	0.8393
rs17271583	12	19334811	1.27E-05	0.8854
rs17506327	13	101942907	2.01E-06	0.8402
rs2581717	18	66234261	5.05E-06	0.8287
rs6099111	20	54370639	1.85E-06	0.5752
rs6626447	X	144741334	1.17E-06	0.2937

4.5 Discussion & Conclusions

ESRF replication was unsuccessful with 439 cases and 400 controls. The power of the replication study is not very high and this may mean that true associations would not be found.

Table 17. Examples of power to detect associations.

These examples are for a disease with 2% prevalence and odds ratio of 1.3 (for the minor allele), and assuming the causal SNP (or a SNP in perfect LD) is tested.

Sample Size (Case/Control)	**Minor Allele Frequency**	**Power**
576/ 3000 (ESRF discovery)	0.05 0.10	0.85 0.76
439/400 (ESRF replication)	0.05 0.10	0.52 0.39

Given the small sample sizes in both cohorts, false positives in the discovery cohort and lack of power to detect effects in the replication are both possible, and it is difficult to say which is more likely without a larger study. In the next chapter I describe the WTCCC3 renal transplant dysfunction study which was carried out as an extension to this study, in order to try to increase power to confirm these associations and find new ones.

None of the 'cloud' SNPs were on the Illumina list of problem SNPs between platforms. Given that the same SNPs are showing these signals in two unrelated phenotypes, and none of them have any other nearby SNPs showing the usual 'tower' of associations from SNPs in LD, we concluded that these are also SNPs with

genotyping problems between the two platforms and reported our findings to Illumina.

5 WTCCC3 Study in Renal Transplantation Dysfunction

In this chapter I describe a genome-wide association study in renal transplant dysfunction (RTD) funded by the Wellcome Trust Case Control Consortium and led by a management team at King's College London. The study was based on the design of the renal transplantation GWAS described in Chapter 4 and includes the samples genotyped for that project, along with further samples from multiple centres and clinical data from NHS Blood and Transplant (NHSBT).

In this chapter I will present several pieces of work I have carried out using the data in this study:

1. Analysis of the primary time-to-failure phenotype, including a preliminary covariates analysis.
2. End-stage renal failure (ESRF) and intracranial haemorrhage (ICH) binary phenotype analysis using logistic regression.
3. Application of SNPHarvester to search for interactions.
4. An adaptation of the SNPHarvester programme which will use the time-to-event data instead of transforming to a binary phenotype.

5.1 Introduction

The aim of this project, as with the smaller King's College London pilot study, was to search for genetic variants that affect the long-term prognosis of a transplanted kidney. The design of this study was slightly different to the design of the pilot study. Stricter inclusion and exclusion criteria were applied, and samples were

contributed by a large consortium consisting of many transplant centres around the United Kingdom and Ireland. Phenotype and covariate data was collected for all UK transplants from NHSBT. Information about the Irish transplant patients was obtained directly from the Irish collaborators. I carried out a comprehensive analysis of all of the data collected on the donors, recipients and transplants to identify covariates that would help to establish the baseline hazard for the survival function. These covariates were then included in the survival model to search for SNPs that were significantly associated with survival time of the graft.

The results of the methods comparison in chapter 3 showed that SNPHarvester was the most successful method for finding the previously identified statistical interaction in psoriasis. I applied this method to the RTD data by transforming the survival phenotype into 3 binary phenotypes - survival at 3 years, 5 years and 10 years.

5.2 Contributions

The work outlined in this chapter is a part of a large collaborative effort. I am one of the analysts on this project, and we have divided the work in some cases, and in others we have duplicated efforts to ensure correctness. In Table 18 I list the work that is presented in this chapter and specify who has done it, with JM referring to myself and CF referring to our collaborator Chris Franklin. In some cases, particularly in the QC, the work was not carried out by me but I feel it is important to present the major findings here in order to understand the data, the analysis and results.

Table 18. Division of analysis work

Work	Analyst(s)
QC of 4 separate cohorts & all file merging	CF
PCA for population stratification	CF, JM
Covariates Analysis	JM
Survival Analysis (all models)	CF, JM
SNPHarvester application and development work	JM

5.3 Genotypes, Phenotypes and Covariates

Genotyping of the majority of the samples for this study was carried out at the Wellcome Trust Sanger Institute on Illumina Quad660 chips. A further set of samples were previously genotyped for the KCL pilot study on Illumina Quad610 chips (see Chapter 4). The transplant centre in Dublin had previously genotyped a cohort of transplant recipients, and we included this data in our study. The matching donors from Dublin were then genotyped as a part of the main cohort at the Sanger Institute. These cohorts will be referred to for the remainder of this chapter as Dublin (Dublin recipients only, genotyped in Dublin), KCL and Sanger (remaining Dublin samples genotyped at the Sanger).

We obtained the phenotype information, along with several clinical measures, from NHSBT on 4th May 2012. The time-to-graft-failure is measured from the date of transplantation and an event is recorded if the patient is lost to follow-up, if their kidney fails or if the study period ends without the failure of the kidney. For our purposes the end of the study period is the date of data extraction by NHSBT, 4th

May 2012. The last recorded event before this time is therefore examined and classified as a failure event or a censored event.

A kidney failure event is defined in a number of ways:

1. Return to dialysis
2. Re-transplantation
3. Return to transplant waiting list

Any transplanted kidneys that were still functioning at the last recorded follow-up date were considered to be censored. In this case the time of last known follow-up was recorded, and the event type was censored.

5.3.1 Inclusion and exclusion criteria

We applied inclusion and exclusion criteria to the selection of samples for genotyping, to maximise the information that could be extracted from the available transplants. These criteria are given in Table 19. Selected data on each transplant was gathered from the transplant centres providing the DNA samples as well as from NHSBT. The data collected from the centres was primarily used to ensure proper identification of donors, recipients and transplants, so the NHSBT clinical data could be matched to the correct samples. Transplant centres which provided DNA also provided us with, at a minimum, NHSBT transplant ID, date of transplant, date of birth of recipient and gender of recipient. Data extracted from NHSBT also included these values, and any instances where the dates and genders did not match up could then be followed up to ensure correct identification of samples. If there was any ambiguity the donor and recipient samples were not genotyped and the transplant event was not included in the analysis.

Table 19. Inclusion and exclusion criteria

Inclusion criteria	Exclusion criteria
Cadaveric donors only Donor and recipient DNA available Transplant date before 1/9/06	Age of recipient < 18 years Recipient had failure event within 3 months of transplant Recipient had no kidney failure event AND the date of the last recorded event is less than three years from transplant date (follow-up of less than 3 years) Any transplant type other than single kidney only

5.3.2 Whole-genome amplification

Some of the DNA samples sent to us for this project did not contain enough DNA for genotyping. In order to increase numbers as much as possible we carried out whole-genome amplification on these samples prior to genotyping. This was also carried out at the Sanger Institute. This set of samples will be referred to as the WGA cohort, and together with the Sanger, KCL and Dublin cohorts make up all of the samples for this study.

5.3.3 Survival-related phenotypes

1. Date of transplant
2. Date of first re-transplant, if occurred
3. Date of first reinstatement on to transplant waiting list, if occurred
4. Date of first transfer back to dialysis, if occurred.
5. Date of last event recorded on the database
6. Last event type: Still living / death
7. gsurv - graft survival, in days
8. gcens - gsurv event type; (1=censored, 2=graft failure event)

The first 6 pieces of information allow us to calculate the survival time, in days, of the graft, or the time to censoring, to check the values for gsurv and gcens that were given to us by NHSBT. Note that we assume that death preceded by dialysis is due to loss of graft function, and this is counted as a graft failure event. Death NOT preceded by dialysis is considered a censored event.

5.3.4 Clinical data

The clinical data available for this study is summarised in the following tables. Table 20 and Table 21 summarise the numeric and categorical variables respectively, and include the number of missing values. Table 22 shows the drug treatments for the recipients at transplantation, 3-months and 12-months after transplantation. Table 23 lists the number of transplants carried out at each centre.

Table 20. Summary of numeric clinical variables

Variable	Mean (range)	Missing
Age of donor at death/transplant	42.69 (1-81)	0
*Age of recipient at transplant	45.62 (16-79)	0
Serum creatinine at 3 months (μmol/l)	157.1 (55-872)	1143
SC 12 months	153.5 (56-992)	998
SC 24 months	156.3 (52-854)	1220
SC 36 months	159.4 (55-934)	1245
SC 48 months	160.1 (51-836)	1427
SC 60 months	158.5 (53-862)	1704
Cold ischaemic time (CIT) in minutes	1151 (0-3180)	1362

***One recipient in the NHSBT data was 16 years old, which is below our threshold of 18. The samples from this transplant were not genotyped and it was not included in any further analyses**

Table 21. Summary of categorical clinical variables

Variable	Factor level counts	Missing
Sex of donor (male/female)	Male: 2039 Female: 1619	1
Sex of recipient (male/female)	Male: 2310 Female: 1349	0

Variable	Factor level counts	Missing
Graft number	1: 3251 2: 362 3: 40 4: 6	0
Graft number	1: 3251 2+: 408	0
Diabetes	Yes: 201 No: 2729	729
Type of transplant	Kidney only: 3630 **En-bloc kidney: 3 **Kidney & pancreas: 20	0
Recipient ethnicity (should be all Caucasian)	Caucasian: 3450 ***Other: 4	205
Recipient blood group	O: 1709 A: 1473 B: 335 AB: 138	4

Variable	Factor level counts	Missing
Donor blood group	O: 1604 A: 1313 B: 229 AB: 91	422
Blood group mismatch	No: 3063 Yes: 176	420
Donor type	Heart beating: 3417 Non-heart beating: 242	0
Intracranial haemorrhage	No: 1260 Yes: 1957	442
Primary renal disease	58 categories with 1-343 in each	1685
HLA mismatches at A,B,DR	27 categories with 4-983 in each	0
Acute rejection 12 months (derived from union of previous)	Yes: 566 No: 1190	1903

****Not genotyped and removed from further analysis as the study is limited to kidney-only transplants**

*****Not genotyped and removed from further analysis as the study is limited to Caucasian recipients**

The type of transplant and recipient ethnicity are included only to check that they are correct. Transplant type should always be kidney-only and recipient ethnicity should always be Caucasian.

Table 22. Frequencies of patients on different drugs at transplant, 6 months and 12 months

Drug and treatment time	No	Yes	Missing
Azathioprine - transplant	1340	981	1338
Azathioprine - 3 months	1403	961	1295
Azathioprine - 12 months	1488	974	1197
ALG induction - transplant	2252	66	1341
Cyclosporin - transplant	995	1329	1335
Cyclosporin - 3 months	1224	1178	1257
Cyclosporin - 12 months	1394	1132	1133
Mycophenolate - transplant	1349	969	1341
Mycophenolate - 3 months	1074	941	1644
Mycophenolate - 12 months	1151	837	1671
Other - transplant	1391	922	1346
Other - 3 months	2008	261	1390
Other - 12 months	1995	347	1317
OKT induction - transplant	2310	8	1341
Prednisolone - transplant	375	1949	1335
Prednisolone - 3 months	340	1818	1501

Prednisolone - 12 months	466	1783	1410
Tacrilomus - transplant	1373	948	1338
Tacrilomus - 3 months	1028	995	1636
Tacrilomus - 12 months	965	1024	1670
Steroids - 3 months	28	193	3438
Steroids - 12 months	26	234	3399

Table 23. Number of transplants carried out at each participating transplant centre

Centre name	**Number of transplants**
Newcastle	184
Leeds	368
Dublin	447
Leicester	114
Nottingham	10
Sheffield	77
Belfast	420
Cambridge	76
London, Royal Free	11
London, Royal London	96
London, Guy's	367

London, St George's	125
Portsmouth	29
Oxford	171
Plymouth	67
Birmingham	227
Coventry	41
Liverpool	42
Manchester	258
Glasgow	159
London, WLRTC	9
Cardiff	361

The tables above show that there was a great deal of missingness in the clinical data. Some of this is due to having less information from certain centres. Table 24 shows missingness for 3 variables by centre. Dublin in particular is missing quite a lot of data for some phenotypes. Dublin is not part of NHSBT and we requested the data from them directly. Our original request for data from Dublin and NHSBT was a reduced set of covariates which were easy to retrieve quickly and were known to be complete and reliable in the NHSBT database. They were also sensible clinical measures that may have an effect on a transplanted kidney. We later made a request for further clinical data but much of this data could not be provided by Dublin. The extra data was sent by NHSBT but the Belfast transplant centre was mistakenly left

out of this additional data set. The inclusion of any of these extended clinical variables in our survival modelling would have resulted in either the removal of all Dublin and Belfast transplants from the study or would have required estimating the missing data. This is often done by using the mean, median or mode depending on the type and distribution of the data. While this is a reasonable approach in some circumstances, it is probably not wise to do this for all of the transplants from a centre. We therefore proceeded with the analysis on the original smaller set of covariates which were available for all transplant centres including Belfast and Dublin.

Table 24. Total transplants by centre and missing data for selected phenotypes

Centre	Total transplants	Missing CIT	Missing diabetes	Missing 3mo creatinine
Newcastle	184	27	0	50
Leeds	368	1	0	65
Dublin	447	311	309	78
Leicester	114	7	0	23
Nottingham	10	0	0	0
Sheffield	77	0	0	4
Belfast	420	420	420	420
Cambridge	76	0	0	3
Royal Free	11	1	0	0
Royal London	96	23	0	32

Centre	Total transplants	Missing CIT	Missing diabetes	Missing 3mo creatinine
Guy's	367	84	0	133
St George's	125	0	0	21
Portsmouth	29	0	0	4
Oxford	171	45	0	65
Plymouth	67	0	0	3
Birmingham	227	115	0	86
Coventry	41	12	0	4
Liverpool	42	0	0	1
Manchester	258	1	0	14
Glasgow	159	143	0	35
WLRTC	9	2	0	0
Cardiff	361	170	0	102

5.4 Genotype Quality Control

Genotype quality control procedures for this study were somewhat more complicated than normal due to genotyping taking place in four different groups. QC was carried out on each of the 4 cohorts individually before genotypes were merged, and one final check was carried out to ensure the genotypes and samples in the merged set passed all QC criteria. Most of the work on the QC was carried out by the lead

analyst at the Sanger Institute (See Table 18), but the results are presented here for completeness.

5.4.1 Quality Control Thresholds and Results

The first step in the QC was to exclude SNPs for low call rates (see Table 25. Initial screening of SNPs with low call rates). This is a fairly generous threshold, and another more strict threshold was applied after sample QC. This was done first, before sample call rates were calculated, so that missingness at very bad SNPs did not push samples with borderline call rates over the threshold for removal.

Table 25. Initial screening of SNPs with low call rates

	Sanger	WGA	KCL	Dublin
Number of samples	3728	990	1201	365
Number of pre-QC SNPs	594398	578723	562831	599011
SNP call rate < 95%	4992	15675	2855	590477

***After merging QCed data sets**

The next stage of QC was to calculate sample (e.g. individual) statistics and remove those not passing our thresholds. These results are given in Table 26. At this stage a PCA was run on each LD-pruned cohort, including HapMap samples from continental ethnic groups JPT, CHB, GIH, YRI and CEU. Samples showing genetic similarity to groups other than the CEU samples were removed.

Table 26. QC results on samples

	Sanger	**WGA**	**KCL**	**Dublin**
Number of samples	3728	990	1201	365
Number of SNPs (after SNP call rate exclusions)	589406	563051	559976	590477
Sample call rate <98%	35	160	37	5
Heterozygosity outliers (3sd)	67 (3 high, 6 low)	22 (11 high, 11 low)	28 (10 high, 18 low)	9 (3 high, 6 low)
IBD>0.1	87	13	167	17
PCA for ethnicity	78	20	54	10
Gender mismatch	18	24	6	1

*After merging QCed data sets

The final stage of QC was to calculate further SNP statistics and remove any further problem SNPs

Table 27. QC results on SNPs

	Sanger	**WGA**	**KCL**	**Dublin**
Number of samples	3501	702	1201	**192
Number of SNPs (after call rate exclusions)	589406	563051	559976	590477
SNP Call rate <98%	7916	25088	3841	13093
HWE p-value < $1x10^{-6}$	9975	9100	2039	5562
MAF < 0.01	29392	27404	18165	24530
Non-random missingness	6642	5813	1376	1132

***After merging QCed data sets**

****333 samples passed sample QC, 192 of these had required phenotype info**

The final numbers of SNPs and samples from each cohort that passed QC, along with the final combined numbers, are given in Table 28. These are the final numbers of SNPs and samples that will be available for analysis.

Table 28. SNPs and samples passing QC from each cohort

Cohort	**SNPs**	**Samples**
Sanger	546055	3501
WGA	522274	702
KCL	538050	1201
Dublin	563011	192
Total	512535	5596

5.4.2 PCA for Identifying Cohort differences

After merging the four cohorts, a PCA without the HapMap data was carried out to identify systematic differences from other sources. Our main aim was to identify and correct for any differences due to genotyping four separate cohorts. We plotted PCA results, colouring the samples by cohort (Figure 14).

PC1 is clearly separating the KCL and Dublin samples, while PC4 shows some separation of the WGA samples from the other three cohorts. PC5 - PC8 do not seem to correlate with cohort labels. PC2 is pulling a small number of KCL and Sanger samples away from the rest, while PC3 spreads some of the Sanger and WGA samples out slightly. PC2 did not separate samples by cohort, but we believed it may be due to some residual population stratification that was not excluded by the thresholds applied to the cohorts during QC before merging. We excluded further samples along this axis at a threshold of -0.1.

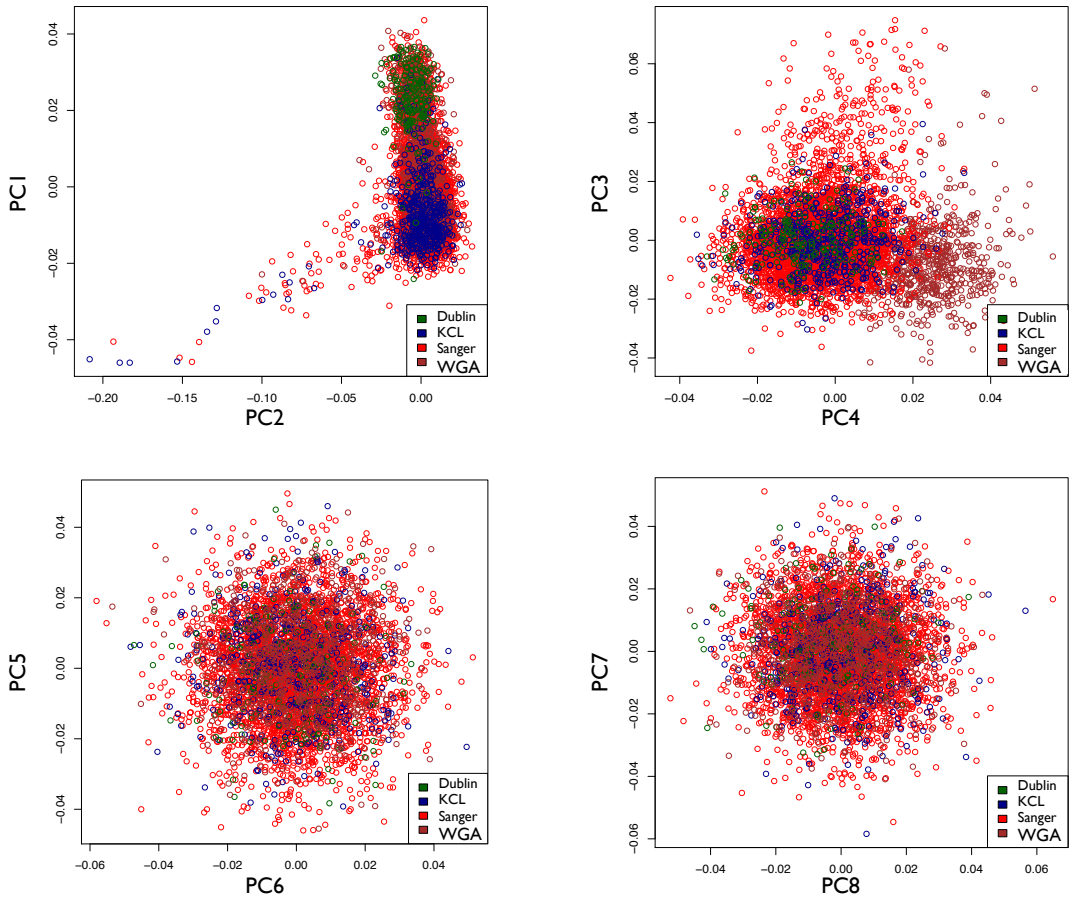

Figure 14. PCA axes 1-8 coloured by cohort.

5.5 Methods

5.5.1 Merging data from WGA samples

Genotype data from the WGA samples had to be merged with the non-WGA samples before analysis. WGA genotypes were first merged with the Sanger cohort, as these had been genotyped at the same location on the same chips, reducing potential sources of variability. This work was carried out by the lead analyst at the Sanger institute, but I present the main findings and the decisions we took as a result.

There were two options for merging the WGA and Sanger cohorts. Intensity data was available for all samples from these cohorts, so it was possible to merge the genotypes at the intensity level and recall the genotypes together. The genotypes were also called independently so it was also possible to merge the data at the genotype level. Merging two different data sets could cause spurious associations or an overall deflation of p-values if there are systematic differences between the cohorts. In order to determine the best way to merge the data it was done both ways, then we carried out pseudo case/control analysis using the WGA cohort as cases and the Sanger cohort as controls. The transformed (-log10(p)) p-values were then examined for a departure from the distribution of expected values. The results shown in Figure 15, indicate a larger inflation of the test statistic when merged at the intensity level. Additionally, the number of samples removed for QC failures was higher when merging intensities and calling the data together. Called separately, a sample missingness threshold of 98% resulted in 34 genomic and 138 WGA samples being removed, for a total of 172 samples. Called together, the same threshold removed 241 samples. In order to retain as many samples as possible and avoid a deflation of p-values due to differences in genotyping, we decided to merge the genotypes rather than merge the intensities and re-call genotypes. All work presented in this thesis, including the previously-presented QC results, is based on the data merged at the genotype level.

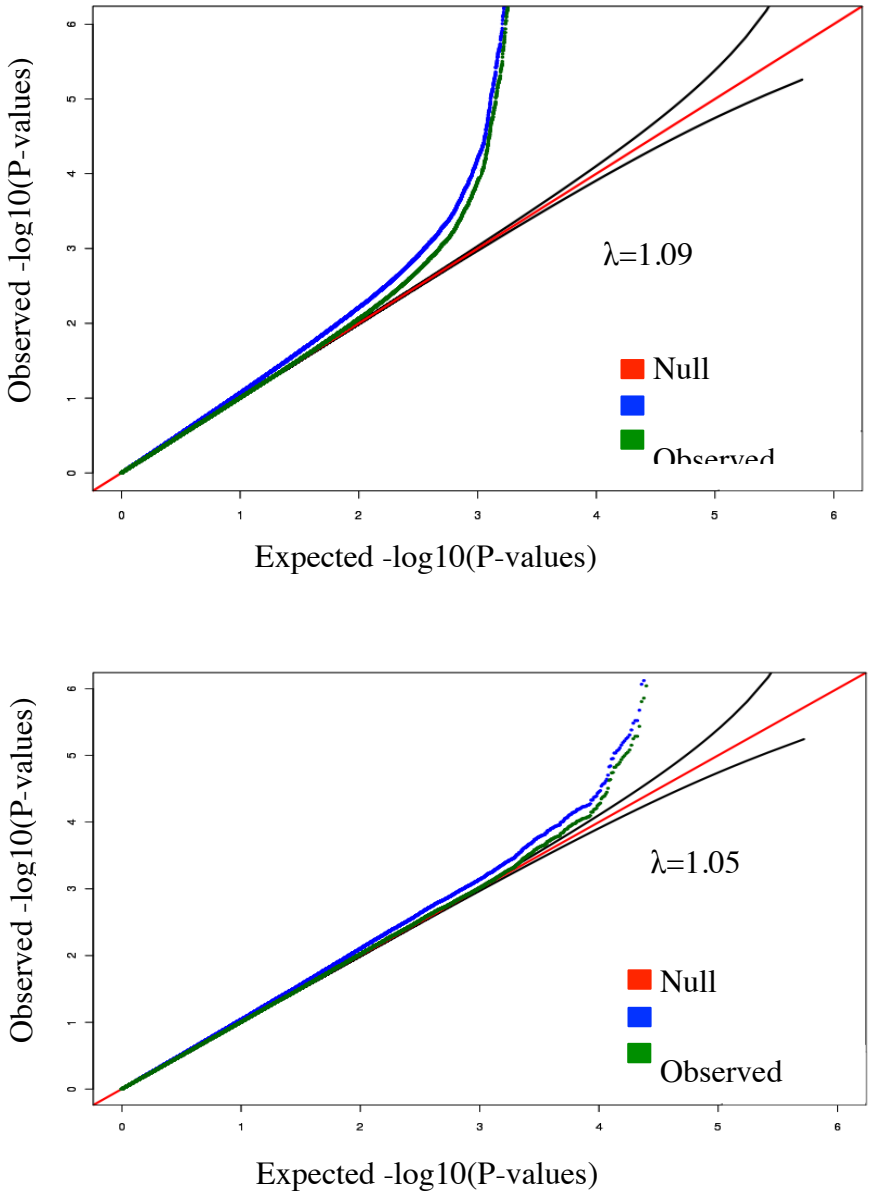

Figure 15. QQ plot from two pseudo-association studies

WGA and Sanger samples after merging data at the intensity level and re-calling genotypes (top) and calling genotypes separately and merging at the genotype level (bottom).

5.5.2 Covariates Analysis

Given the high level of missingness of some of the clinical data, we included only covariates that were complete or nearly complete in the analysis of covariates (see Clinical data). In some cases, clinical data that may be important but was left out of the model due to missing data can be checked on after the analysis; for example any hits in the survival analysis can be checked against SNPs known to be associated with diabetes or acute rejection, to see if there is any overlap.

The final form of the analysis of the covariates is given in Equation 13.

Equation 13. Model as fit to covariates using Cox Proportional Hazards modelling.

$$h_i(t) = \exp(\alpha(t) + Recip\_ct + Graftno + Rage + Dage + Rsex + Dsex)$$

where Recip_ct is the transplant centre where the recipient received the kidney, Graftno is the number of transplants the recipient has received (including present), Rage is recipient age, Dage is donor age, Rsex is recipient sex and Dsex is donor sex.

Kaplan-Myer plots for covariates were also generated to be examined for any groupings of factors.

Results for the survival analysis of covariates are presented in a table, each row of which shows the coefficient ß for each numeric variable or single level of a factor, the exponential of ß, its standard error, a z-score and a p-value for the z-score. Taking the exponential of the coefficient gives the relative risk (RR). For a factor this is the risk relative to a reference group, with the reference group having a hazard of 1. If exp(ß) is higher than one, this is an increase in the hazard, and the interpretation for a factor is that the group is at increased risk relative to the reference group, and is more likely to fail sooner. For example if exp(ß) is 1.5, transplants from that group are 1.5 times more likely to fail at any given time *t* than transplants

from the reference group. For the transplant centre the reference group is A1313, which is the centre with the median survival time. Recipients on their first transplant are the reference group for the graft number covariate (GRAFT_NO), and all others (on their 2nd, 3rd and 4th transplant) are grouped together into the group "2+". Dsex and Rsex represent the change in survival time for females (group 2) relative to males.

For numeric variables the exponential of the coefficient represents the effect on survival time for every increase of 1 in that variable. For example, with donor age, exp(ß) represents the increase in risk of failure for each increase in age of one year. For example, if exp(ß) is 1.05, a transplant with a donor of age 30 at time of death is 5% more likely to fail at any time than a transplant from a donor who is 29 at the time of death.

All survival models were fit using the software package R version 2.15.0 (R Development Core Team, 2011), using the survival package version 2.36-1 (Therneau and Lumley, 2011).

5.5.3 Survival Analysis of Genotypes

I carried out a survival analysis using the genetic data by fitting a Cox Proportional Hazards model as I did with the covariates analysis. For this analysis I want to test, for each SNP, a recipient effect, a donor effect, and an interaction effect, all at the same SNP. A full Cox model for a single SNP is shown in Equation 14.

Equation 14. Full Cox Proportional Model for a single SNP

$$h_i(t) = \exp\left(\alpha(t) + \boldsymbol{\beta_0} COV + \beta_1 SNP1R + \beta_2 SNP1D + \beta_3 SNP1R * SNP1D\right)$$

where t is time in days, $\alpha(t)$ is the baseline hazard function (incorporating the intercept), $\boldsymbol{\beta_0}$ is the set of coefficients for the covariates, β_1 is the effect of the recipient genotype, β_2 is the effect of the donor genotype, and β_3 is the interaction effect between donor and recipient genotype. In this model the unit of analysis is the transplant, so in the model above, the locus for SNP1R and SNP1D is the same, but the people contributing genotypes for each are different; they are the donor and the recipient contributing to the same transplant.

When designing the study every attempt was made to identify transplants in which both donor and recipient DNA was available along with the required phenotypes and covariates. In some cases the transplant event was missing a donor or a recipient, due to individuals removed for quality control issues. If donor and recipient effects with all covariates were fit and tested in the same model, then only the intersection of samples with donor and recipient genotypes could be used for all effect estimations. In order not to waste data, the models were fit separately for each effect so that, for example, a transplant event whose donor was not genotyped would not be eliminated when estimating the recipient effect. In Equation 15 below, Model 1 uses all samples with recipient genotypes available and calculates coefficients and p-values for recipient effects at all SNPs. Model 2 uses all samples with donor genotypes available and calculates coefficients and p-values for all SNPs. Model 3 includes only samples with both donor and recipient genotypes available, and again calculates coefficients and p-values for all SNPs.

Equation 15. Single-SNP regression for transplant data showing the a) recipient-only model, b) donor-only model and c) full interaction model

$$a)\ h_i(t) = \exp\left(\alpha(t) + \boldsymbol{\beta_0} COV + \beta_1 SNP1R\right)$$

$$b)\ h_i(t) = \exp\left(\alpha(t) + \boldsymbol{\beta_0} COV + \beta_2 SNP1D\right)$$

$$c)\ h_i(t) = \exp\left(\alpha(t) + \boldsymbol{\beta_0} COV + \beta_1 SNP1R + \beta_1 SNP1D + \beta_3 SNP1R * SNP1D\right)$$

The significance of the term of interest is then estimated by dropping the term to be tested and carrying out a 1 degree of freedom likelihood ratio test between the two models. For models one (recipient) and two (donor), the SNP coefficient is dropped to leave only the baseline hazard and the covariates. In model 3 (full interaction model) the interaction term is dropped, leaving the baseline hazard, covariates and main effects in the model.

5.5.4 Application of SNPHarvester to binary survival phenotype

In its published form, SNPHarvester runs on case/control phenotypes only (see Chapter 2). In order to apply SNPHarvester to the data from this renal transplant dysfunction study, I transformed the data to 3 different binary phenotypes; 3, 5, and 10-year survival. For each of these, any graft failure events occurring before the specified time threshold were considered cases, and any grafts that definitely survived until that time were controls. Any transplants that had a censored event time before the time threshold were dropped from the analysis, as it was not possible to allocate them with certainty to the case or control group. This produced three case/control data sets, each of which had a different number of cases, controls, and total samples (see Table 29). The three-year survival phenotype had the largest number of total samples but the smallest number of cases. As time moves on, more

failure and censoring events occur, which results in a larger number of cases but a smaller total number of samples.

Table 29. Binary survival phenotype case/control counts

Phenotype	Cases	Controls	Total
3-year survival	155	1920	2075
5-year survival	268	1585	1853
10-year survival	449	540	989

I ran SNPHarvester on these data sets with the stopping criteria numsuccessiverun=5. This means that if the algorithm creates a path through the entire data set 5 times without any pairs of SNPs exceeding the p-value threshold, the process will terminate.

5.6 Results

5.6.1 Covariates Analysis

The table of results from the survival analysis on the covariates is given in Table 30. For this analysis there were 961 failure events from a total of 3658 transplants.

Table 30. Covariates analysis results

Variable/Level	ß	exp(ß)	se(ß)	z	Pr(>|z|)
RECIP_CT/B6313	-0.062	0.940	0.195	-0.319	0.749
RECIP_CT/BH	0.436	1.55	0.171	2.6	0.011*
RECIP_CT/C0301	0.037	1.04	0.251	0.148	0.882

RECIP_CT/C0851	-0.432	0.649	1.01	-0.426	0.670
RECIP_CT/C1201	-0.188	0.829	0.351	-0.535	0.592
RECIP_CT/CH	0.663	1.94	0.172	3.842	1.22E-04***
RECIP_CT/D0101	-0.261	0.771	0.386	-0.676	0.499
RECIP_CT/F0708	-0.304	0.738	1.01	-0.300	0.764
RECIP_CT/F1212	0.091	1.096	0.275	0.332	0.740
RECIP_CT/G1501	0.179	1.196	0.186	0.965	0.335
RECIP_CT/H1305	-0.596	0.551	0.326	-1.827	0.068.
RECIP_CT/J2102	0.414	1.51	0.385	1.08	0.282
RECIP_CT/K4102	-0.307	0.736	0.242	-1.267	0.205
RECIP_CT/L3395	-0.017	0.983	0.367	-0.047	0.963
RECIP_CT/M1202	-0.187	0.830	0.208	-0.897	0.370
RECIP_CT/M1701	0.258	1.29	0.316	0.816	0.414
RECIP_CT/N2117	-0.031	0.969	0.436	-0.071	0.943
RECIP_CT/P1101	0.076	1.08	0.203	0.373	0.709
RECIP_CT/SG516	0.115	1.12	0.202	0.568	0.570
RECIP_CT/T0701	0.232	1.26	0.724	0.321	0.748
RECIP_CT/W7001	0.059	1.06	0.182	0.325	0.745
GRAFT_NO/2+	0.452	1.57	0.098	4.61	4.1E-06***
RAGE	-0.003	0.997	0.003	-1.30	0.192

DAGE	0.020	1.02	0.002	8.75	<2e-16***
RSEX/2	0.086	1.09	0.066	1.29	0.197
DSEX/2	-0.068	0.934	0.066	-1.04	0.300

Signif. codes: 0 '*' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1**

Although there are several clinical variables significantly associated with graft failure times, the overall variance explained is only 4.3%.

Transplant centres BH and CH are significantly different from the median A1313 (not shown, RR=1). Both of these have a higher RR than A1313. H1305 has a lower RR with a marginal p-value of 0.07. A Kaplan-Meier plot of survival time by centre is given in Figure 15. Recipients who have received at least one previous graft have a significantly higher RR (Figure 16), as do recipients who receive kidneys from older donors (Figure 17). Donor and recipient sex have no effect on the time to graft failure (Figure 19). Based on these results we decided to include centre as a 3-group factor, with the groups associated with low, average and high survival times. BH & CH form the low group, H1305 is the high group, and the remainder of the centres are in a single 'average' group. Another factor of two levels indicating if it is the recipient's first transplant was also included, along with donor and recipient age at transplant as a numerical covariate. We included recipient age because although it is not associated with survival time in this analysis (Figure 18) other studies have found it to be important (Summers et al., 2010).

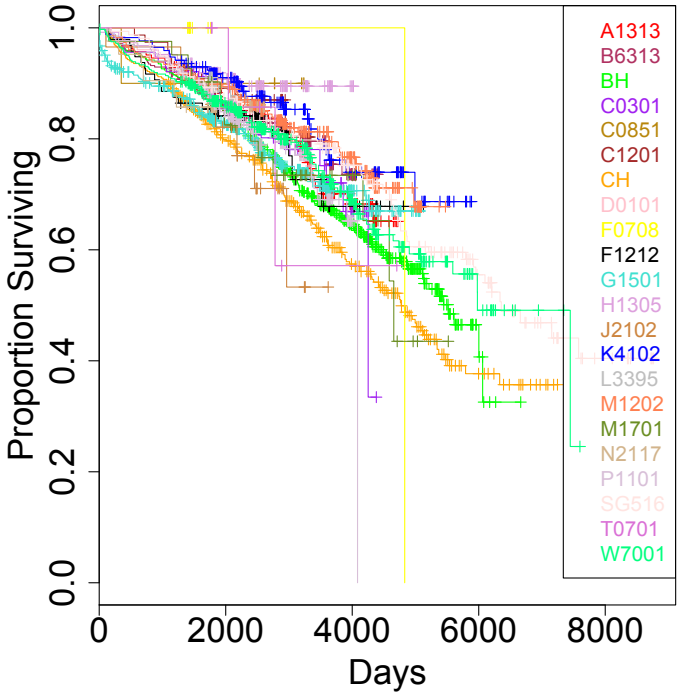

Figure 16. Kaplan-Meier plot of survival time by transplant centre.

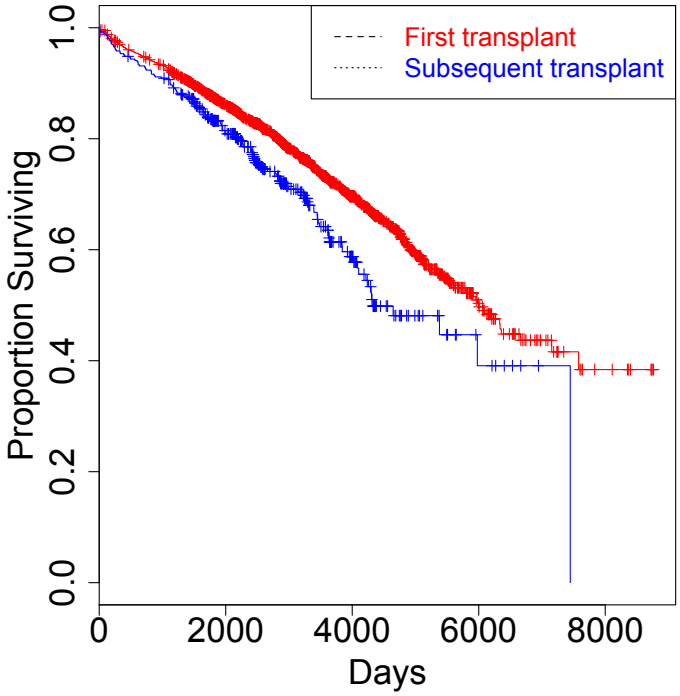

Figure 17. Kaplan-Meier plot of survival time by transplant number.

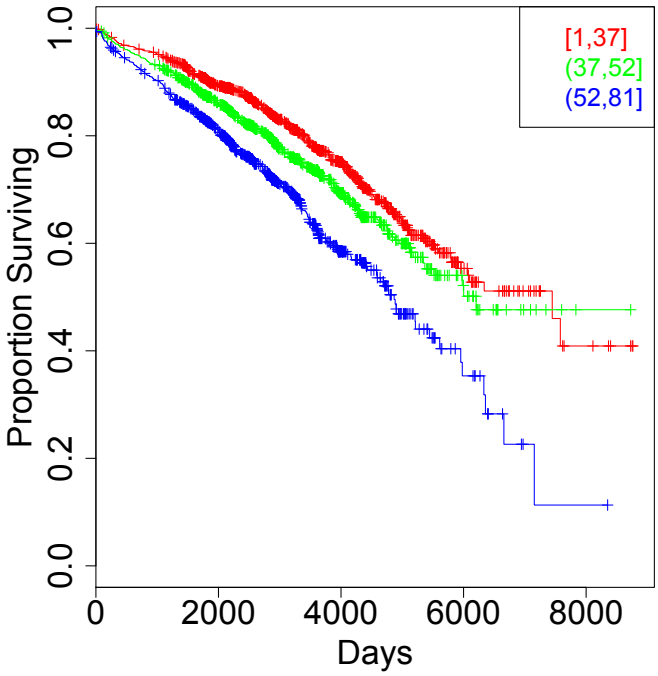

Figure 18. Kaplan-Meier plot of survival time by donor age, divided into three equal-sized groups.

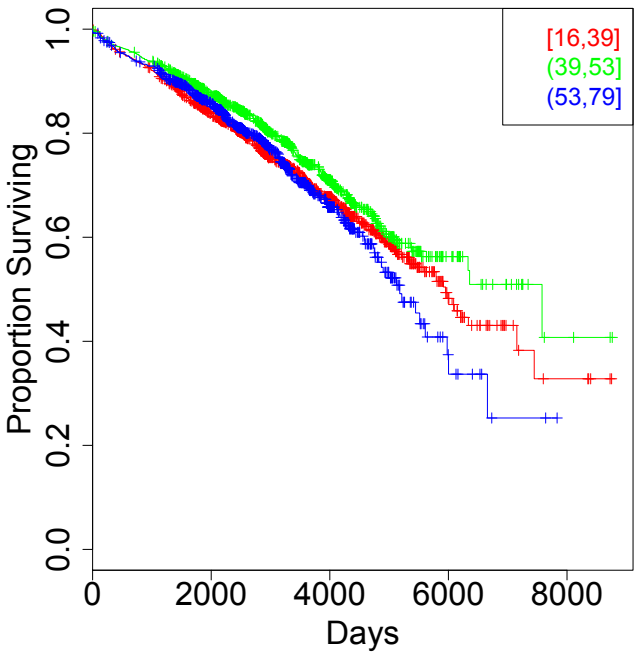

Figure 19. Kaplan-Meier plot of survival time by recipient age, divided into three equal-sized groups.

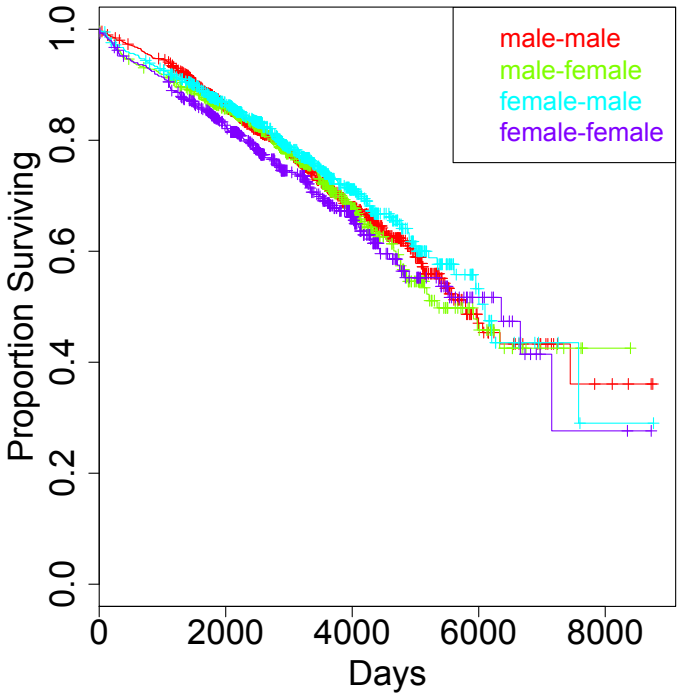

Figure 20. Kaplan-Meier plot of survival time by gender (donor-recipient).

5.6.2 Survival Analysis

The results of fitting the Cox proportional hazards model to the time-to-graft-failure phenotype with covariates (3-group transplant centre, donor age, recipient age, binary first transplant indicator and first 4 PC axes) are shown in Manhattan plots in Figure 21. Over all of the results for all three effects, there were no genome-wide significant results.

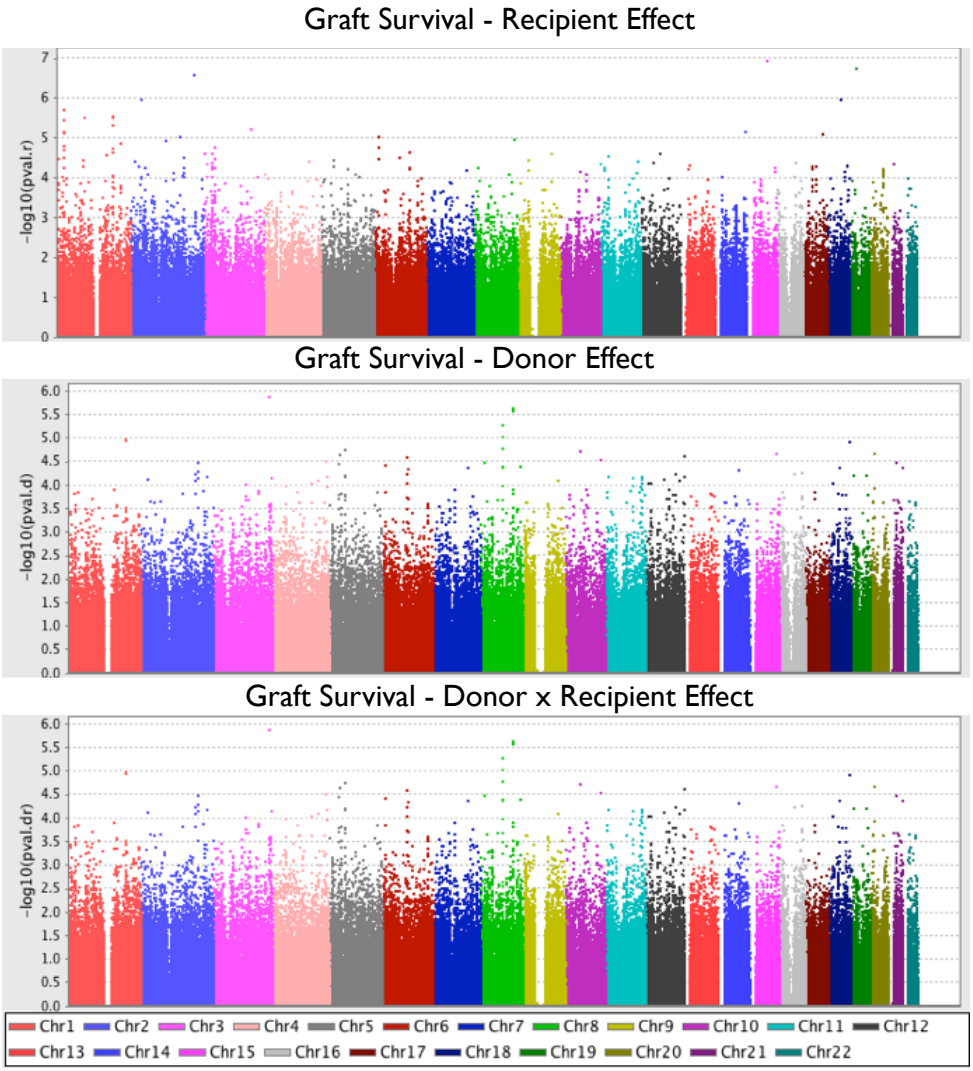

Figure 21. Manhattan plot of results for survival analysis of graft failure phenotype

Table 31 Top SNP associations from survival analysis of graft failure

Chr	SNP	P-value	n	Effect
15	rs11072010	1.19×10^{-7}	2686	R
19	rs7250100	1.78×10^{-9}	2690	R
2	rs1465804	2.54×10^{-7}	2686	R
3	rs4859121	1.37×10^{-6}	2202	D
8	rs4734914	2.39×10^{-6}	2204	D
8	rs7827856	2.73×10^{-6}	2202	D
3	rs922784	5.22×10^{-6}	2204	D
8	rs7002197	9.78×10^{-6}	2203	D
3	rs4859121	1.75×10^{-5}	2202	I

5.6.3 SNPHarvester

The first run of SNPHarvester was on the 10-year survival phenotype, which included 449 cases and 540 controls. This analysis did not produce any significant single SNP or SNP pair results. Given the small of samples in the analysis, the power to detect any associations would be very low. Additionally, there are no significant findings in the survival analysis, so it is unlikely that SNPHarvester would have found single-SNP associations that the survival analysis could not. The five-year survival phenotype included 268 cases and 1585 controls. The only finding suggested by SNPHarvester was a single SNP, rs3754949 (recipient), on

chromosome 2. This SNP had a recipient p-value of 0.551 in our single-SNP survival analysis.

The three-year survival phenotype included 155 cases and 1920 controls for analysis. SNPHarvester selected 1558 SNP pairs with scores that exceeded the Bonferroni-corrected threshold. I examined the distribution of the chi-square test statistic to see if there was any evidence of strongly association pairs of SNPs. This is a left-truncated distribution because the information about the SNP pairs with test statistics lower than the left-hand threshold was not saved, and so is not plotted. There are some SNPs at the top end of the distribution which are slightly separated from the group and this could possibly represent a set of strong interaction signals.

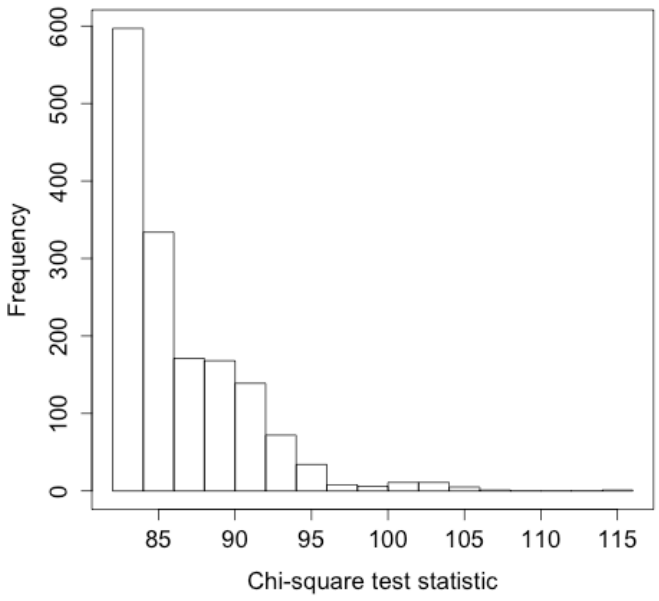

Figure 22. Histogram of test statistics for SNP pairs selected by SNPHarvester. All test statistics are chi-square distributed with 8 degrees of freedom.

The test underlying the SNPHarvester results is a chi-square test. It is sensible to check for rare genotype combinations between the significant pairs, as it is not recommended to apply a chi-square test to contingency tables with cell values of less

than five. A screen of the SNPHarvester results revealed that only 44 of the 1558 SNP pair contingency tables had all cells with a value of more than five. The remainder of the results must be regarded as untrustworthy as the chi-square test would not be appropriate. Furthermore, 24 of the remaining 44 pairs involved at least one SNP that was a known copy number polymorphism (Dayem Ullah et al, 2012). This would likely lead to genotype calling problems. The remaining 20 SNP pairs are given in Table 32. None of these SNPs are previously published GWAS associations, and none of them are non-synonomous coding variants.

Table 32. Location of Inteactions discovered by SNPHarvester

X^2(2d.f.)	SNP1	Chr 1	Position 1	SNP 2	Chr 2	Position 2
89.2	rs1888861	1	156797961	rs9868633	3	169741291
90.37	rs1888861	1	156797961	*rs9846780	3	187095439
89.96	rs1888861	1	156797961	*rs1568302	3	187100520
91.7	rs1888861	1	156797961	*rs1518872	3	187121796
84.14	rs1888861	1	156797961	*rs13065558	3	187122532
83.26	rs1888861	1	156797961	rs709081	3	191446862
82.74	rs1888861	1	156797961	rs1923200	6	23651526
83.8	rs1888861	1	156797961	rs310409	6	81798503
85.32	rs9870672	3	57404880	rs3926405	2	106563295
92.46	rs9870672	3	57404880	#rs7582470	2	121575818
83.36	rs9870672	3	57404880	#rs895491	2	121578827
85.37	rs9870672	3	57404880	rs4557160	3	164747930
82.95	rs9870672	3	57404880	rs6449012	4	14317116

X^2(2d.f.)	SNP1	Chr 1	Position 1	SNP 2	Chr 2	Position 2
84.17	rs9870672	3	57404880	rs9352851	6	81392952
83.65	rs9870672	3	57404880	rs960874	10	10491092
88.45	rs9870672	3	57404880	[+]rs1924507	10	10521475
88.96	rs9870672	3	57404880	[+]rs71003	10	119349761
89.06	rs9870672	3	57404880	rs4441549	20	6042852
83.87	rs6461942	7	26485860	rs11786417	8	133135790
82.73	rs7185918	16	85901065	rs987611	18	1306068

***Pairwise LD between these SNPs ranges from 0.85-0.98**

[#]Pairwise LD is 0.96

[+]Pairwise LD is 0.98

There are only 4 SNPs in the 'SNP1' column of Table 32. Two of these appear in several rows, paired with several SNPs that are in high LD, indicating that the same association interaction signal is being picked up. This further reduces the total number of signals detected to 15 interactions.

5.7 Extension of SNPHarvester to Survival Data

SNPHarvester in its published format takes binary phenotype data. In order to make the best use of the data available in the renal transplant dysfunction study, I have worked on an extension to SNPHarvester to allow it to take time-to-event data. At the heart of the original SNPHarvester is the idea of using a score test to quickly estimate the association between a pair of SNPs and a phenotype. In the original

SNPHarvester as used on case/control data, a chi-square test is the appropriate score test to apply as a fast, easily computed hypothesis test for binary data. For a Cox Proportional Hazards model, the appropriate score test is a log-rank test statistic (Peto and Peto, 1972). I therefore decided to implement the log-rank test as a function in SNPHarvester to make the best use of the time-to-event data.

A further limitation to the original SNPHarvester is that it reads in case/control phenotype data. In order to apply a log-rank test I also need SNPHarvester to read in and store both the event time and the event type. This will require an extension of the current function to read in the case/control data.

5.7.1 Log-rank test

The log-rank test takes time-to-event data for two or more groups and compares their survival functions for equivalence. The goal of this test is similar to testing groups for equivalence using a t-test, or a non-parametric alternative such as rank-sum test. The concept can be visualized by imagining a Kaplan-Meier plot with multiple lines representing survivor functions for different groups, and there is separation between some or all of the lines. The statistical question is then identifying whether or not these survival functions are truly different - do the subjects in different groups have significantly different risks? A standard test for differences between groups would not take into account the existence of censored data, and when this is present the log-rank test or similar tests with different weighting options are an appropriate alternative. For the remainder of this explanation of the logrank test, I will use the terms 'survive' and 'die' to describe the possible outcomes in a survival analysis, rather than the term 'failure' or 'graft failure', which is more appropriate to the renal transplant dysfunction study data.

For the log-rank test, each observed time point can be considered separately, and we can create a contingency table separating individuals by their survival status and stratifying them by the grouping variable of interest (see Table 33).

Table 33. Counts of at risk subjects at a single time point t(i) stratified by group and event type

Event	Group 1	Group 2	Group 3	Total
Died	d_{i1}	d_{i2}	d_{i3}	d_i
Not Died	n_{i1}-d_{i1}	n_{i2}-d_{i2}	n_{i3}-d_{i3}	n_i-d_i
At risk	n_{i1}	n_{i2}	n_{i3}	n_i

In this table, the total number of subjects at risk at time t_i is denoted by n_i, and these are divided into K groups (K=3 for this example). The groups could be different treatment groups or, for an example using genetic information, it could be people carrying one, two or three risk alleles at a SNP. The total number who died at this time point is denoted by d_i, divided into d_{i1} (group 1), d_{i2} (group 2) and d_{i2} (group 3). The number not dying at a time point is the difference between the total number at risk and the number who died, which will include any censored events at this time point. As we consider subsequent time points the number at risk is affected by the number of previous deaths, but it is also affected by the number of censored events. Subjects who have been censored are no longer at risk, and will not be included in the 'at risk' value for subsequent time points. In this way, subjects who are eventually censored still contribute to the calculation of the risk at time points before the censoring event, making full use of all the available data.

Once the contingency table is created, its contribution to the test statistic is calculated in a familiar manner, based on the differences between the observed and expected scores and dividing by the variance. The general form for the test statistic is given in matrix notation in Equation 16.

Equation 16. Test statistic for the logrank test with multiple groups

$$Q = \left[\sum_{i=1}^{m} W_i(d_i - \hat{e}_i)\right]' \left[\sum_{i=1}^{m} W_i \hat{V}_i W_i\right]^{-1} \left[\sum_{i=1}^{m} W_i(d_i - \hat{e}_i)\right]$$

where d_i is a vector containing the number of deaths in each of K-1 groups at time i, $\hat{e}$ is the expected number of deaths in the same groups at the same time point, and $(d_i - \hat{e}_i)$ is the difference between these two vectors. Any K-1 of the K groups can be used, but one must be left out to avoid the centre matrix on the right-hand side of Equation 16 being singular. W_i is a vector of length m of weights assigned to the each time point. For a basic log-rank test, this weight is always 1. Tests related to the log-rank test such as the modified Peto-Prentice, generalized Wilcoxon and Tarone and Ware apply different weights to the model at different time points. To calculate the covariance matrix, we can consider that each time point in the log-rank test is a trial with a certain number of successes (deaths) from the available subjects (number at risk). Since the number at risk is reduced at each time point, this is a series of trials *without replacement*, and the hypergeometric distribution is appropriate. The covariance matrix for the test statistic for multiple groups is therefore derived from the hypergeometric distribution. $\hat{V}_i$ in Equation 16 is the covariance matrix for d_i, of size K-1 by K, and the formulae for the within-group variances in the diagonal elements (Equation 17) and between-group covariances in the off-diagonal elements (Equation 18) are given below.

Equation 17. Diagonal (variance) elements of the covariance matrix of d_i

$$\hat{V}_{kki} = \frac{n_{ki}(n_i - n_{ki})d_i(n_i - d_i)}{n_i^2(n_i - 1)}, k = 1,2, \dots, K-1$$

Equation 18. Off-diagonal (covariance) elements of the covariance matrix of d_i

$$\hat{V}_{kli} = \frac{n_{ki}n_{li}d_i(n_i - d_i)}{n_i^2(n_i - 1)}, k, l = 1,2, \dots, K-1, k \neq l$$

In our specific case of two interacting SNPs we have 9 possible groups of people with different 2-locus genotypes, so our contingency table will have 9 groups, and the covariance matrix will be 8 by 9 elements.

5.7.2 Implementation

The implementation of the logrank test extension to SNPHarvester will involve changes to the original code in Java, and a call to R from within the Java program to carry out the logrank test. As a reminder, SNPHarvester is a hill-climbing algorithm that involves calculating multiple paths through the set of SNPs from random starting points (see Chapter 2). The path through the SNPs is determined by the score of the current set of SNPs and the potential replacement SNP, and any increase in score will result in a new SNP joining the group and an old SNP being dropped. The algorithm describing a single hill-climbing path through the data is called PathSeeker. Pathseeker is called repeatedly by SNPHarvester to create multiple paths through the data from random starting points. An outline of the PathSeeker algorithm taken from the SNPHarvester paper (Sun et al., 2009) is given in Figure 23. The changes to be made to this algorithm are:

1. The data passed to the PathSeeker algorithm will be a single set of genotypes with associated time-to-event phenotype data - an event time and event type (censored or failure).
2. The Score test called by PathSeeker (whenever Score() is seen in the algorithm) will be the logrank test instead of the current chi-square test.

Reading in the time-to-event data will involve creating a new read-data function. This function will create arrays to hold the time (in days) and the censoring status (0=event, 1=censored) for every response, then read this data in and store it in the time and censor arrays.

The implementation of the log-rank test in SNPHarvester will involve using rJava (Urbanek, 2011), an interface between Java and R. The R package 'Interval' (Fay and Shaw, 2010, Schwarz et al., 2010) contains a function called *ictest*. This function will perform a log-rank test on multiple groups. An example of the structure of the data to be tested for a pair of SNPs is in Table 34.

Table 34. Time to event data for the log-rank test of two SNPs.

This table represents the data at a single time point *i*. There will be *M* such tables built and the test statistics are summed over all of them.

	Died	**Not Died**	**At risk**
AABB	d_{i1}	$n_{i1}-d_{i1}$	n_{i1}
AABb	d_{i2}	$n_{i2}-d_{i2}$	n_{i2}
AAbb	d_{i3}	$n_{i3}-d_{i3}$	n_{i3}
AaBB	d_{i4}	$n_{i4}-d_{i4}$	n_{i4}
AaBb	d_{i5}	$n_{i5}-d_{i5}$	n_{i5}
Aabb	d_{i6}	$n_{i6}-d_{i6}$	n_{i6}
AABB	d_{i7}	$n_{i7}-d_{i7}$	n_{i7}
AABb	d_{i8}	$n_{i8}-d_{i8}$	n_{i8}
AAbb	d_{i9}	$n_{i9}-d_{i9}$	n_{i9}
Total	d_i	n_i-d_i	n_i

The results from SNPHarvester will be the same as when the chi-square score test is used. The program will return a list of SNP pairs whose scores exceed the threshold for significance. The interaction between these SNP pairs is not necessarily significant. As with the original SNPHarvester, the scores represent the combined effect of the two SNPs, and these still need to be tested for interactions. This can then be done in a post-processing step by fitting a Cox proportional hazards model containing the two SNPs of interest and their interaction.

Algorithm 1. PathSeeker algorithm

Notation: q: iteration number.
Input:
D: a dataset of N_d case samples and N_u control samples genotyped at L SNP makers.
k: k-SNP groups.
T: statistical significance threshold.
Output:
M: a collection of k-SNP groups which pass the statistical testing.
E: the local optimum (i.e. the active set at the end of a path).

/* Phase 1-Initialization */
Randomly select k SNPs to form an active set A

/* Phase 2-Iteration */
Initialize $q=0$ and $IsSwap \leftarrow$ True
while $IsSwap$=True **do**
  $IsSwap \leftarrow$ False
  for $i=1$ to L **do**
    if SNP_i does not belong to A **then**
      $h \leftarrow \arg\max_{j=1,\ldots,k} Score(A+\{SNP_i\}-\{SNP_{s_j}\})$
      if $Score(A+\{SNP_i\}-\{SNP_{s_h}\}) > Score(A)$ **then**
        Update A as $A \leftarrow A+\{SNP_i\}-\{SNP_{s_h}\}$
        $IsSwap \leftarrow$ True
        /* Harvest the significant interacting SNPs */
        if $Score(A) > T$ **then**
          Put A into M
        end if
      end if
    end if
  end for
  $q++$ **/*Record iteration number */**
end while
$E \leftarrow A$ **/*Record the local optimum after convergence */**
return M, E

Figure 23. SNPHarvester's PathSeeker algorithm from Yang et al, showing the use of the *Score* function

5.7.3 Application to Transplant Data

The ultimate aim of adding a time-to-event option to SNPHarvester is to apply it to the renal transplant dysfunction GWAS data. The current solution of transforming

time to event data to a binary variable using a threshold is not optimal, since any data that is censored before the threshold time is not used at all. In application to the renal data, SNPHarvester did not find any associations in the 5-year and 10-year thresholds, even though there were more failure events. This is because as time passes there are also more censored events, so the number of 'controls' in the binary data drops substantially, resulting in a loss of power. The log rank test, making use of the censored data and all of the failure events, may increase power enough to detect further interactions.

5.8 Discussion and conclusions

The survival analysis of the time to graft failure did not lead to convincing results. It is possible that there are no SNPs that are directly related to post-transplant prognosis. This is an observational study with many unmeasured environmental factors that could explain the variability in graft failure times, such as drug compliance, drug response, smoking status, diet and exercise. These may account for a great deal of the variability seen in graft survival times, and genetics may play a minor role. It is also possible that we lack power to find associations with the number of transplants studied here. Immunosuppressive drugs are changing and possibly improving, and these drugs may be masking genetic effects. Given that there were no strong single-SNP associations it isn't surprising that I have not found any significant interactions. Our initial hypothesis, that there could be interactions between genetic variants in donors and recipients, is exactly the problem which immunosuppressive drugs are meant to overcome.

HLA type is known to be an important factor in the success of kidney transplantation. It seems plausible that SNPs in HLA genes may be associated with

the survival time of a kidney transplant. There are several factors which can explain why they did not show any association in this study. First, many transplants are matched on HLA type to minimise the risk of rejection. This will reduce the number of mismatches and make any effects more difficult to detect. Secondly, immunosuppressive drugs may mask the effects of mismatches. Finally, SNPs do not uniquely tag the HLA types that have been found to influence prognosis in transplantation, so a single-SNP analysis may not pick up on any effect the HLA types may have on survival time.

At the time of writing imputation of SNP data up to 1000 genomes coverage was underway. Analysis of imputed SNPs may lead to stronger single-SNP association signals or may confirm that our current top associations are not very strong. Once the imputation analysis is complete, replication of the most likely candidate SNPs will be carried out in a European collection of kidney transplant donors and recipients (Opelz, 1992).

The list of interactions suggested by SNPHarvester from the 3-year survival phenotype was extensive, but can be reduced drastically by eliminating SNP pairs with rare combined genotypes. The results can be reduced further still by eliminating SNPs that are known copy-number polymorphisms, as these may lead to poor genotype calling. The remaining pairs represent 15 interaction signals involving 19 SNPs, which is a sensible set of choices to take forward for replication in another cohort of transplant patients. The extension to SNPHarvester, making full use of the time-to-event data, may produce further interactions for follow-up.

6 Discussion and Conclusions

The overarching theme to this thesis is the search for evidence of interactions between genetic variants which are associated with clinical phenotypes. Each chapter presented a piece of work aimed at understanding how, and indeed if, statistical analysis can be used to identify these interactions.

The usual approach in a genome-wide association study is to start with genetic variants and apply a statistical method, using associations between SNPs and a phenotype to suggest SNPs, genes, pathways or proteins that may be disrupted in disease states. In Chapter 2 I turned this around and started with a real association between a pair of SNPs and psoriasis. I tested methods designed to search for interactions in GWAS data to assess their performance on real GWAS data. Many such software packages are tested on simulated data to demonstrate their effectiveness. Simulations are carried out to create samples and genotypes that are representative of patterns which are hypothesised to be associated with diseases or traits in real samples from populations of interest. This is a useful technique but can be misleading, as the simulated data is often generated using the same assumptions that were made in designing the software, and the software is very likely to perform well. These assumptions are a necessary starting point, but they may not reflect reality. It would be far easier to design software if the designer knew what to search for, but of course this isn't known until it is found. For this study, I starting with a real sample, with real phenotype information, that had a previously-identified and replicated interaction between two SNPs. My aim was to test software on real data with all its complications. If the software packages had been able to find the real interaction, then there would be some evidence that these approaches are sensible.

Although applying SNPHarvester followed by the HyperLasso did identify the SNP pair in the psoriasis study, it would not have been the most likely candidate for follow-up had this truly been a search for novel interactions. With the lack of a good replication cohort, it is impossible to say if any of the other more likely candidates might be real. Sparse Partitioning, did not select the previously identified interaction. Asking Sparse Partitioning to search for a large number of predictors slows it down considerably. Given the large number of strong, single-SNP associations in this data set it is unlikely that the interaction effect would be strong enough to be selected in the model. However Sparse Partitioning did successfully select several independent signals which have since proved to be true associations. Finally, Random Jungle was also unsuccessful at selecting the interacting pair. In the analysis for this study I ran the program as recommended by the authors. It is possible that different settings for Random Jungle, such as limiting tree depth, could change the results.

In Chapter 3 I searched for an overall enrichment in low p-values for interactions between pairs of SNPs suggested by published protein-protein interactions. My approach of starting from the protein-protein interactions (PPIs) could again be considered the reverse of a standard association study. The PPIs that were my starting point have generally been identified in laboratory experiments, and have not necessarily been shown to occur in the human body. The proteins may not be expressed in the same tissue, in which case there is no chance for them to interact. They may be expressed in the same tissue but be used in other biological processes so that they do not come into contact with each other. However it is reasonable to hypothesize that at least some of these PPIs do happen in *vivo*, and may be disrupted in disease states. Finding an enrichment of statistical interaction signals in these

SNPs would have suggested that biological interactions in vivo can be found using statistical analysis of SNP data. However given the rarity of confirmed, convincing examples in the literature, and the absence of the enrichment I was hoping to find in the PPI study, it is still not clear that this approach is possible or practical.

Chapter 4 outlined a genome-wide association study in renal transplant dysfunction that was a pilot study for the larger project outlined in chapter 5, and eventually these two projects were combined. In both studies the main phenotype was time to graft failure, necessitating the use of survival modelling. Of particular interest in this study was the interaction between donor and recipient genomes, which was assessed by including an interaction term between donor and recipient genotypes at the same SNP. Searching between different SNPs, both within and across the two genomes, was carried out by the application of SNPHarvester to a binary conversion of the survival times. Although this has produced a list of suggested SNP pairs, these require further investigation and replication before there can be confidence in any findings.

Finally, an extension to SNPHarvester to handle time-to-event data was presented. The log-rank test, as the score test for the Cox proportional hazards model, provides a fast test of association for survival data without fitting the full model. This is being implemented as an alternative function in SNPHarvester. At the time of writing I was still working on this software, with the intention of applying it to the transplant study data when it is complete. This will avoid the necessity of converting the survival times to a binary phenotype, which reduces power through the binary reclassification and through the loss of information from cases censored before the binary decision threshold.

Throughout this thesis the hypothesis that SNP-SNP interactions can be found in GWAS studies through statistical methods has been tested in several ways, using many different methods. None of these methods has proven successful at finding novel interactions or at providing evidence that these interactions exist. This search is certainly not definitive but it does suggest that, as others have found, these interactions are not easy to find. If they do exist in association with human diseases and traits they almost certainly have small effect sizes, making them very difficult to identify.

7 References

1000 GENOMES PROJECT CONSORTIUM 2010. A map of human genome variation from population-scale sequencing. *Nature,* 467**,** 1061-73.

ANDERSON, C. A., PETTERSSON, F. H., CLARKE, G. M., CARDON, L. R., MORRIS, A. P. & ZONDERVAN, K. T. 2010. Data quality control in genetic case-control association studies. *Nature protocols,* 5**,** 1564-1573.

BATESON 1909. Discussion on the Influence of Heredity on Disease, with special Reference to Tuberculosis, Cancer, and Diseases of the Nervous System: Introductory Address. *Proceedings of the Royal Society of Medicine,* 2**,** 22-30.

BRITISH KIDNEY PATIENT ASSOCIATION 2012a. Acute Renal Failure.

BRITISH KIDNEY PATIENT ASSOCIATION 2012b. Chronic Kidney Disease.

BRITISH KIDNEY PATIENT ASSOCIATION 2012c. Haemodialysis.

BRITISH KIDNEY PATIENT ASSOCIATION 2012d. Peritoneal Dialysis.

CORDELL, H. 2002. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics*.

CORDELL, H. 2009. Genome-wide association studies: Detecting gene-gene interactions that underlie human diseases. *Nature reviews Genetics*.

COX, D. 1972. Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological),* 34**,** 187-220.

CRICK, F. 1970. Central dogma of molecular biology. *Nature,* 227**,** 561-563.

CRICK, F. H. 1958. On protein synthesis. *Symposia of the Society for Experimental Biology,* 12**,** 138-163.

ABU Z DAYEM ULLAH, NICHOLAS R LEMOINE AND CLAUDE CHELALA, SNPnexus: a web server for functional annotation of novel and publicly known genetic variants (2012 update), Nucleic Acids Research, 2012, 40(W1):W65-W70.

EVANS, D. M., SPENCER, C. C. A., POINTON, J. J., SU, Z., HARVEY, D., KOCHAN, G., OPPERMANN, U., OPPERMAN, U., DILTHEY, A., PIRINEN, M., STONE, M. A., APPLETON, L., MOUTSIANAS, L., MOUTSIANIS, L., LESLIE, S., WORDSWORTH, T., KENNA, T. J., KARADERI, T., THOMAS, G. P., WARD, M. M., WEISMAN, M. H., FARRAR, C., BRADBURY, L. A., DANOY, P., INMAN, R. D., MAKSYMOWYCH, W., GLADMAN, D., RAHMAN, P., SPARCC, S. R. C. O. C., MORGAN, A., MARZO-ORTEGA, H., BOWNESS, P., GAFFNEY, K., GASTON, J. S. H., SMITH, M., BRUGES-ARMAS, J., COUTO, A.-R., SORRENTINO, R., PALADINI, F., FERREIRA, M. A., XU, H., LIU, Y., JIANG, L., LOPEZ-LARREA, C., DÍAZ-PEÑA, R., LÓPEZ-VÁZQUEZ, A., ZAYATS, T., BAND, G., BELLENGUEZ, C., BLACKBURN, H., BLACKWELL, J. M., BRAMON, E., BUMPSTEAD, S. J., CASAS, J. P., CORVIN, A., CRADDOCK, N., DELOUKAS, P., DRONOV, S., DUNCANSON, A., EDKINS, S., FREEMAN, C., GILLMAN, M., GRAY, E., GWILLIAM, R., HAMMOND, N., HUNT, S. E., JANKOWSKI, J., JAYAKUMAR, A., LANGFORD, C., LIDDLE, J., MARKUS, H. S., MATHEW, C. G., MCCANN, O. T., MCCARTHY, M. I., PALMER, C. N. A., PELTONEN, L., PLOMIN, R., POTTER, S. C., RAUTANEN, A., RAVINDRARAJAH, R., RICKETTS, M., SAMANI, N., SAWCER, S. J., STRANGE, A., TREMBATH, R. C., VISWANATHAN, A. C., WALLER, M., WESTON, P., WHITTAKER, P., WIDAA, S., WOOD, N. W., MCVEAN, G., REVEILLE, J. D., WORDSWORTH, B. P., BROWN, M. A., DONNELLY, P., TASC, A.-A.-A. S. C. & WTCCC2, W. T. C. C. C. 2011. Interaction between ERAP1 and HLA-B27 in ankylosing spondylitis implicates peptide handling in the mechanism for HLA-B27 in disease susceptibility. *Nature Genetics,* 43**,** 761-767.

FAY, M. P. & SHAW, P. A. 2010. Exact and Asymptotic Weighted Logrank Tests for Interval Censored Data: The interval R Package. *Journal of Statistical Software,* 36**,** 1-34.

FISHER, R. 1918. Fisher: The correlation of relatives on the assumption of Mendeian inheritance. *Proc Roy Soc Edin*.

FISHER, R. A. 1925. *Statistical Methods for Research Workers,* Edinburgh, Oliver and Boyd.

GENOVESE, G., FRIEDMAN, D. J., ROSS, M. D., LECORDIER, L., UZUREAU, P., FREEDMAN, B. I., BOWDEN, D. W., LANGEFELD, C. D., OLEKSYK, T. K., USCINSKI KNOB, A. L., BERNHARDY, A. J., HICKS, P. J., NELSON, G. W., VANHOLLEBEKE, B., WINKLER, C. A., KOPP, J. B., PAYS, E. & POLLAK, M. R. 2010. Association of trypanolytic ApoL1 variants with kidney disease in African Americans. *Science (New York, N.Y.),* 329**,** 841-845.

LUCIA A HINDORFF, PRAVEEN SETHUPATHY, HEATHER A JUNKINS, ERIN M RAMOS, JAYASHRI P MEHTA, FRANCIS S COLLINS, AND TERI A MANOLIO. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. (2009).*106*(23), 9362–9367.

HOGGART, C. J., WHITTAKER, J. C., DE IORIO, M. & BALDING, D. J. 2008. Simultaneous analysis of all SNPs - supplementary. *PLoS Genet,* 4**,** e1000130.

HOWIE, B. N., DONNELLY, P. & MARCHINI, J. 2009. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet,* 5**,** e1000529.

INTERNATIONAL HAPMAP CONSORTIUM 2003. The International HapMap Project. *Nature,* 426**,** 789-796.

IOANNIDIS, J. P. A. 2005. Why most published research findings are false. *PLoS medicine,* 2**,** e124.

JOHNSON, R. J., FUGGLE, S. V., O&APOS;NEILL, J., START, S., BRADLEY, J. A., FORSYTHE, J. L. R., RUDGE, C. J. & TRANSPLANT, K. A. G. O. N. B. A. 2010. Factors influencing outcome after deceased heart beating donor kidney

transplantation in the United Kingdom: an evidence base for a new national kidney allocation policy. *Transplantation,* 89**,** 379-386.

KNIGHT, J., SPAIN, S. L., CAPON, F., HAYDAY, A., NESTLE, F. O., CLOP, A., CONSORTIUM, W. T. C. C., CONSORTIUM, G. A. O. P., CONSORTIUM, I.-C. F. P., BARKER, J. N., WEALE, M. E. & TREMBATH, R. C. 2012. Conditional analysis identifies three novel major histocompatibility complex loci associated with psoriasis. *Human Molecular Genetics*.

KOCIERZ, M., SIEKIERA, U., KOLONKO, A., KARKOSZKA, H., CHUDEK, J., CIERPKA, L. & WIĘCEK, A. 2011. -174G/C interleukin-6 gene polymorphism and the risk of transplanted kidney failure or graft loss during a 5-year follow-up period. *Tissue antigens,* 77**,** 283-290.

B T LEE, V KUMAR, T A WILLIAMS, R ABDI, A BERNHARDY, C DYER, S CONTE, G GENOVESE, M D ROSS, D J FRIEDMAN, R GASTON, E MILFORD, M R POLLAK, AND A CHANDRAKER. 2012. The APOL1 Genotype of African American Kidney Transplant Recipients Does Not Impact 5-Year Allograft Survival. *American journal of transplantation : official journal of the American Society of Transplantation and the American Society of Transplant Surgeons*.

LANDER, E. S., LINTON, L. M., BIRREN, B., NUSBAUM, C., ZODY, M. C., BALDWIN, J., et al. 2001. Initial sequencing and analysis of the human genome. *Nature*, *409*, 860–921.

LEHNE, B. & SCHLITT, T. 2009. Protein-protein interaction databases: keeping up with growing interactomes. *Human genomics,* 3**,** 291-297.

MAHER, B. 2008. Personal genomes: The case of the missing heritability. *Nature*, Nov 06, p.7218.

MANTEL, N. 1967. The detection of disease clustering and a generalized regression approach. *Cancer Research,* 27**,** 209-220.

MARCHINI, J., DONNELLY, P. & CARDON, L. 2005. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nature Genetics,* 37**,** 413-417.

MELLS, G. F., FLOYD, J. A. B., MORLEY, K. I., CORDELL, H. J., FRANKLIN, C. S., SHIN, S.-Y., HENEGHAN, M. A., NEUBERGER, J. M., DONALDSON, P. T., DAY, D. B., DUCKER, S. J., MURIITHI, A. W., WHEATER, E. F., HAMMOND, C. J., DAWWAS, M. F., CONSORTIUM, U. P., 3, W. T. C. C. C., JONES, D. E., PELTONEN, L., ALEXANDER, G. J., SANDFORD, R. N. & ANDERSON, C. A. 2011. Genome-wide association study identifies 12 new susceptibility loci for primary biliary cirrhosis. *Nature Genetics,* 43**,** 329-332.

MOTSINGER-REIF, A. A., REIF, D. M., FANELLI, T. J. & RITCHIE, M. D. 2008. A comparison of analytical methods for genetic association studies. *Genetic Epidemiology,* 32**,** 767-778.

MÜLLER-STEINHARDT, M., EBEL, B. & HÄRTEL, C. 2007. The impact of interleukin-6 promoter -597/-572/-174genotype on interleukin-6 production after lipopolysaccharide stimulation. *Clinical and experimental immunology,* 147**,** 339-345.

MÜLLER-STEINHARDT, M., FRICKE, L., MÜLLER, B., EBEL, B., KIRCHNER, H. & HÄRTEL, C. 2004. Cooperative influence of the interleukin-6 promoter polymorphisms -597, -572 and -174 on long-term kidney allograft survival. *American journal of transplantation : official journal of the American Society of Transplantation and the American Society of Transplant Surgeons,* 4**,** 402-406.

NAIR, R. P., DUFFIN, K. C., HELMS, C., DING, J., STUART, P. E., GOLDGAR, D., GUDJONSSON, J. E., LI, Y., TEJASVI, T., FENG, B.-J., RUETHER, A., SCHREIBER, S., WEICHENTHAL, M., GLADMAN, D., RAHMAN, P., SCHRODI, S. J., PRAHALAD, S., GUTHERY, S. L., FISCHER, J., LIAO, W., KWOK, P.-Y., MENTER, A., LATHROP, G. M., WISE, C. A., BEGOVICH, A. B., VOORHEES, J. J., ELDER, J. T., KRUEGER, G. G., BOWCOCK, A. M., ABECASIS, G. R. & COLLABORATIVE ASSOCIATION STUDY OF

PSORIASIS 2009. Genome-wide scan reveals association of psoriasis with IL-23 and NF-kappaB pathways. *Nat Genet,* 41**,** 199-204.

OPELZ G, MYTILINEOS J, WUJCIAK T, SCHWARZ V, BACK D. 1972. Asymptotically efficient rank invariant procedures. *Journal of the Royal Statistical Society, Series A,* 135**,** 185-207.

PETO, R. & PETO, J. 1972. Asymptotically efficient rank invariant procedures. *Journal of the Royal Statistical Society, Series A,* 135**,** 185-207.

PRICE, A. L., PATTERSON, N. J., PLENGE, R. M., WEINBLATT, M. E., SHADICK, N. A. & REICH, D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics,* 38**,** 904-909.

PURCELL, S., NEALE, B., TODD-BROWN, K., THOMAS, L., FERREIRA, M. A. R., BENDER, D., MALLER, J., SKLAR, P., DE BAKKER, P. I. W., DALY, M. J. & SHAM, P. C. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics,* 81**,** 559-575.

R DEVELOPMENT CORE TEAM 2011. A language and environment for statistical computing. . 2.15.0 ed. Vienna, Austria: R Foundation for Statistical Computing.

REIMHERR, M. & NICOLAE, D. L. 2011. You've gotta be lucky: Coverage and the elusive gene-gene interaction. *Annals of human genetics,* 75**,** 105-111.

SÁNCHEZ-VELASCO, P., RODRIGO, E., FERNÁNDEZ-FRESNEDO, G., OCEJO-VINYALS, J. G., RUIZ, J. C., ARNAU, A., LEYVA-COBIÁN, F. & ARIAS, M. 2010. Influence of interleukin-6 promoter polymorphism -174 g/c on kidney graft outcome. *Transplantation proceedings,* 42**,** 2854-2855.

SCHULTE, F., SCHNÜLLE, P., BUGERT, P., KLÜTER, H. & MÜLLER-STEINHARDT, M. 2011. The interleukin-6 promoter (-597/-572/-174)genotype does not affect interleukin-6 production in hemodialysis patients. *Journal of*

interferon & cytokine research : the official journal of the International Society for Interferon and Cytokine Research, 31**,** 639-642.

SCHWARZ, D. F., KÖNIG, I. R. & ZIEGLER, A. 2010. On safari to Random Jungle: a fast implementation of Random Forests for high-dimensional data. *Bioinformatics (Oxford, England),* 26**,** 1752-1758.

SPEED, D. 2010. *Exploring Nonlinear Regression Methods, with Application to Association Studies*. Doctor of Philosophty, University of Cambridge.

STRANGE, A., CAPON, F., SPENCER, C. C. A., KNIGHT, J., WEALE, M. E., ALLEN, M. H., BARTON, A., BAND, G., BELLENGUEZ, C., BERGBOER, J. G. M., BLACKWELL, J. M., BRAMON, E., BUMPSTEAD, S. J., CASAS, J. P., CORK, M. J., CORVIN, A., DELOUKAS, P., DILTHEY, A., DUNCANSON, A., EDKINS, S., ESTIVILL, X., FITZGERALD, O., FREEMAN, C., GIARDINA, E., GRAY, E., HOFER, A., HÜFFMEIER, U., HUNT, S. E., IRVINE, A. D., JANKOWSKI, J., KIRBY, B., LANGFORD, C., LASCORZ, J., LEMAN, J., LESLIE, S., MALLBRIS, L., MARKUS, H. S., MATHEW, C. G., MCLEAN, W. H. I., MCMANUS, R., MÖSSNER, R., MOUTSIANAS, L., NALUAI, A. T., NESTLE, F. O., NOVELLI, G., ONOUFRIADIS, A., PALMER, C. N. A., PERRICONE, C., PIRINEN, M., PLOMIN, R., POTTER, S. C., PUJOL, R. M., RAUTANEN, A., RIVEIRA-MUNOZ, E., RYAN, A. W., SALMHOFER, W., SAMUELSSON, L., SAWCER, S. J., SCHALKWIJK, J., SMITH, C. H., STÅHLE, M., SU, Z., TAZI-AHNINI, R., TRAUPE, H., VISWANATHAN, A. C., WARREN, R. B., WEGER, W., WOLK, K., WOOD, N., WORTHINGTON, J., YOUNG, H. S., ZEEUWEN, P. L. J. M., HAYDAY, A., BURDEN, A. D., GRIFFITHS, C. E. M., KERE, J., REIS, A., MCVEAN, G., EVANS, D. M., BROWN, M. A., BARKER, J. N., PELTONEN, L., DONNELLY, P., TREMBATH, R. C. & GENETIC ANALYSIS OF PSORIASIS CONSORTIUM & THE WELLCOME TRUST CASE CONTROL, C. 2010. A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between HLA-C and ERAP1. *Nature Genetics,* 42**,** 985-990.

SUMMERS, D. M., JOHNSON, R. J., ALLEN, J., FUGGLE, S. V., COLLETT, D., WATSON, C. J. & BRADLEY, J. A. 2010. Analysis of factors that affect outcome after transplantation of kidneys donated after cardiac death in the UK: a cohort study. *Lancet,* 376**,** 1303-1311.

SUN, M., LI, N., DONG, W., CHEN, Z., LIU, Q., XU, Y., HE, G., SHI, Y., LI, X., HAO, J., LUO, Y., SHANG, D., LV, D., MA, F., ZHANG, D., HUA, R., LU, C., WEN, Y., CAO, L., IRVINE, A. D., MCLEAN, W. H. I., DONG, Q., WANG, M.-R., YU, J., HE, L., LO, W. H. Y. & ZHANG, X. 2009. Copy-number mutations on chromosome 17q24.2-q24.3 in congenital generalized hypertrichosis terminalis with or without gingival hyperplasia. *American journal of human genetics,* 84**,** 807-813.

THERNEAU, T. & LUMLEY, T. 2011. Survival: Survival analysis, including penalised likelihood. 2.36-5 ed.: R Foundation for Statistical Computing.

TSOI, L. C., SPAIN, S. L., KNIGHT, J., ELLINGHAUS, E., STUART, P. E., CAPON, F., DING, J., LI, Y., TEJASVI, T., GUDJONSSON, J. E., KANG, H. M., ALLEN, M. H., MCMANUS, R., NOVELLI, G., SAMUELSSON, L., SCHALKWIJK, J., STÅHLE, M., BURDEN, A. D., SMITH, C. H., CORK, M. J., ESTIVILL, X., BOWCOCK, A. M., KRUEGER, G. G., WEGE, W., WORTHINGTON, J., TAZI-AHNINI, R., NESTLE, F. O., HAYDAY, A., HOFFMANN, P., WINKELMANN, J., WIJMENGA, C., LANGFORD, C., EDKINS, S., ANDREWS, R., BLACKBURN, H., STRANGE, A., BAND, G., PEARSON, R. D., VUKCEVIC, D., SPENCER, C. C., DELOUKAS, P., MROWIETZ, U., SCHREIBER, S., WEIDINGER, S., KOKS, S., KINGO, K., ESKO, T., METSPALU, A., LIM, H. W., VOORHEES, J. J., WEICHENTHAL, M., WICHMANN, H. E., CHANDRAN, V., ROSEN, C. F., RAHMAN, P., GLADMAN, D. D., GRIFFITHS, C. E., REIS, A., KERE, J., COLLABORATIVE ASSOCIATION STUDY OF PSORIASIS, GENETIC ANALYSIS OF PSORIASIS CONSORTIUM, PSORIASIS ASSOCIATION GENETICS EXTENSION, WELLCOME TRUST CASE CONTROL CONSORTIUM, NAIR, R. P., FRANKE, A., BARKER, J. N., ABECASIS, G. R., ELDER, J. T. & TREMBATH, R. C. 2012. Identification of 15 new psoriasis susceptibility loci highlights the role of innate immunity. *Nature Genetics,* in press.

TZUR, S., ROSSET, S., SHEMER, R., YUDKOVSKY, G., SELIG, S., TAREKEGN, A., BEKELE, E., BRADMAN, N., WASSER, W. G., BEHAR, D. M. & SKORECKI, K. 2010. Missense mutations in the APOL1 gene are highly associated with end stage kidney disease risk previously attributed to the MYH9 gene. *Human genetics,* 128**,** 345-350.

URBANEK, S. 2011. rJava: Low-level R to Java interface. 0.9-3 ed.

WAHRMANN, M., DÖHLER, B., RUHENSTROTH, A., HASLACHER, H., PERKMANN, T., EXNER, M., REES, A. J. & BÖHMIG, G. A. 2011. Genotypic diversity of complement component C4 does not predict kidney transplant outcome. *Journal of the American Society of Nephrology,* 22**,** 367-376.

WATSON, C. J. E., JOHNSON, R. J., BIRCH, R., COLLETT, D. & BRADLEY, J. A. 2012. A simplified donor risk index for predicting outcome after deceased donor kidney transplantation. *Transplantation,* 93**,** 314-318.

WEALE, M. E. 2010. Quality Control for Genome-Wide Association Studies

Genetic Variation. 628**,** 341-372.

WELLCOME TRUST CASE CONTROL CONSORTIUM 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature,* 447**,** 661-678.

WINKELMAYER, W. C., SUNDER-PLASSMANN, G., HUBER, A. & FÖDINGER, M. 2004. Patterns of co-occurrence of three single nucleotide polymorphisms of the 5,10-methylenetetrahydrofolate reductase gene in kidney transplant recipients. *European journal of clinical investigation,* 34**,** 613-618.

WOLFE, R. A., ASHBY, V. B., MILFORD, E. L., OJO, A. O., ETTENGER, R. E., AGODOA, L. Y., HELD, P. J. & PORT, F. K. 1999. Comparison of mortality in all patients on dialysis, patients on dialysis awaiting transplantation, and recipients of a first cadaveric transplant. *New England Journal of Medicine,* 341**,** 1725-1730.

ZHANG, Y., & LIU, J. S. 2007. Bayesian inference of epistatic interactions in case-control studies. *Nature Genetics*, *39*, 1167–1173.

ZUK, O., HECHTER, E., SUNYAEV, S. R. & LANDER, E. S. 2012. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences of the United States of America,* 109**,** 1193-1198.